Quantification and Mechanistic Analysis of Plant Genome Editing Outcomes using Nanopore Sequencing

A DISSERTATION

SUBMITTED TO THE FACULTY OF

THE UNIVERSITY OF MINNESOTA

BY

Paul Allen Parker Atkins

IN PARTIAL FULFILLMENT OF THE REQUIREMENTS

FOR THE DEGREE OF

DOCTOR OF PHILOSOPHY

Adviser, Daniel F. Voytas

August 2020

Acknowledgements

I am extremely grateful to my adviser, Dan Voytas, for giving me the time, resources, and support necessary to produce this work.

Special thanks to Colby Starker, who's advice and assistance has proven invaluable numerous times, and to Maria Elena Gamo for ensuring the continuation of this work in an unprecedented environment.

Additional thanks to all the members and visiting scholars of the Voytas lab over the years, particularly Tomas Cermak for molding my perspective on molecular biology and genome engineering, Eva Konečná for being an incredible role model labmate, and Michael Maher, Ryan Nasti, Matt Zinselmeier, Evan Ellison, Redeat Tibebu, and James Chamness for countless stimulating conversations and support over the years.

Dedication

For Theodore, whose arrival made this possible.

Synopsis

Precise genome modification via homologous recombination, or gene targeting (GT), allows crop genomes to be tailored to any application or environment. While GT's potential is immense, it tends to be inefficient and technically challenging in plants. These problems are compounded by the slow and low-throughput nature of plant transformation, drastically hindering optimization. More insidiously, these issues result in dependence upon proxies and reporter readouts for estimating GT frequencies that vary between groups and delivery platform making it difficult to compare experimental outcomes.

To enable widespread optimization of plant GT, a universal platform for directly measuring genome editing outcomes at the molecular level that accommodates plant-specific technical constraints is urgently needed. Here I develop such a platform, an amplicon-based analysis pipeline using Oxford Nanopore Sequencing (ONS). ONS has several valuable qualities for a plant GT optimization pipeline, namely its accessibility, speed, and read length, making it feasible for even the smallest labs to perform on-demand sequencing with their own equipment. These strengths are accompanied by a major shortcoming – sequencing error. I mitigate this problem using several approaches in a novel bioinformatics pipeline to minimize the effect of ONS error on estimates of targeted mutagenesis and virtually eliminating its effect on estimates of GT frequencies.

Using this pipeline, I observed a significant impact of both geminiviral replicons (GVRs) and donor sequence divergence on gene targeting frequencies. Additionally, I was able to observe the conversion tracts of hundreds of gene targeting events, revealing their deposition by multiple DNA repair pathways and the prevalence of extremely short tracts, which will inform future optimization efforts. This work establishes a universal pipeline for quantifying plant gene targeting events, facilitating future optimization and communication of results between disparate experimental systems within the plant community.

Table of Contents

List of Tables

List of Figures

CHAPTER ONE

Overcoming bottlenecks in plant gene editing

Paul A.P. Atkins and Daniel F. Voytas

Current Opinions in Plant Biology, 2020

https://doi.org/10.1016/j.pbi.2020.01.002

Preface

Agriculture has reached a technological inflection point. The development of novel gene editing tools and methods for their delivery to plant cells promises to increase genome malleability and transform plant biology. Whereas gene editing is capable of making a myriad of DNA sequence modifications, its widespread adoption has been hindered by a number of factors, particularly inefficiencies in creating precise DNA sequence modifications and ineffective methods for delivering gene editing reagents to plant cells. Here, we briefly overview the principles of plant genome editing and highlight a subset of the most recent advances that promise to overcome current limitations.

Introduction

Gene editing has been made possible by the advent of efficient programmable DNA binding domains, giving researchers the ability to deliver DNA-modifying proteins to any DNA sequence of interest in complex genomes (Chandrasegaran and Carroll 2016). Traditionally, most genome editing approaches require a targeted DNA double-strand break (DSB) at the DNA sequence to be edited (Chandrasegaran and Carroll 2016). Prior to Cas9, protein-based DNA binding domains (e.g. zinc finger or TALEs) were fused to a nuclease domain to create sequence-specific nucleases (SSNs) capable of making targeted DSBs (Chandrasegaran and Carroll 2016). Cas9 possesses innate nuclease activity, and its DNA binding is directed by RNA-DNA base-pairing, greatly simplifying the process of generating a novel SSN (Figure 1-1A) (Gasiunas et al. 2012; Jinek et al. 2012; Cho et al. 2013; Mali et al. 2013). Additionally, Cas9 can be converted to a nickase or simply a DNA-binding domain by inactivation of one or both of its nuclease domains, respectively (Gasiunas et al. 2012; Jinek et al. 2012; Cong et al. 2013).

Targeted Mutagenesis

Targeting of a DSB or SSB (single-strand break) recruits DNA repair machinery that is capable of generating a spectrum of DNA-sequence modifications, each with distinct applications (Figure 1-1) (Chandrasegaran and Carroll 2016; Schmidt, Pacher, and Puchta 2019a). Perhaps the simplest mutagenic DNA repair outcome is targeted mutagenesis, or the creation of single nucleotide polymorphisms (SNPs) or small indels (insertions and/or deletions) at a particular sequence. Indels often create frameshifts, rendering the gene product non-functional and potentially targeting it for silencing through nonsense-mediated decay. These inactivating mutations typically result from either non-homologous end-joining (NHEJ) or microhomology-mediated end-joining (MMEJ). NHEJ frequently results in perfect repair by rejoining blunt ends (normally undetectable), but occasionally bases are removed (or more rarely added) prior to ligation, and small indels are created (Figure 1-1B). Recent work indicates that MMEJ is responsible for many indels formed after a nuclease-

induced DSB (Kregten et al. 2016; Zelensky et al. 2017; Mara et al. 2019). In one form of MMEJ, broken DNA ends are resected a short distance allowing free ssDNA to anneal after which Pol Theta (*TEBICHI* in Arabidopsis) fills in the remaining single-stranded region (Figure 1-1C illustrates one of many possible MMEJ outcomes) (Inagaki et al. 2006). The presence of microhomology at many deletions suggests that MMEJ is a predominant mutagenic repair pathway in both genome editing and genome evolution (Schendel et al. 2015).

While NHEJ and MMEJ are typically thought of as imprecise and mutagenic, they may be utilized to create precise DNA insertions with little or no homology (Maresca et al. 2013; K. Suzuki et al. 2016; Orlando et al. 2010; Nakade et al. 2014). In this approach, linear DNA fragments are inserted into DSB sites. This has most recently been practiced in plants as Intron Targeting (Figure 1-1D) (J. Li et al. 2016; Xu et al. 2019). This approach replaces exons by targeting nucleases to adjacent introns for excision and replacement by a co-delivered fragment. Alternatively, Intron Targeting can be used to insert a splice acceptor and the remainder of the coding sequence into an intron using a single nuclease. Introns targeted in this fashion may be flanked by indels, but since the cut sites are within introns, such mutations can be accommodated. That said, recent analysis of large targeted deletions and inversions in Arabidopsis suggest this may not be necessary due to the high frequency of perfect DSB repair (Schmidt, Pacher, and Puchta 2019b).

Targeting indels using nucleases has many applications beyond simple gene inactivation. In tomato, mutagenesis of promoters controlling key agronomic traits yielded novel alleles and meaningfully altered quantitative traits (Rodríguez-Leal et al. 2017). Additionally, multiplexed targeted mutagenesis has been used to explore *de novo* domestication in tomato and ground cherry (Zsögön et al. 2017; Lemmon et al. 2018; T. Li et al. 2018; Zsögön et al. 2018). Large deletions and inversions have been created by targeting multiple DSBs to distant sites, (Qi et al. 2013; T. Čermák et al. 2017; Ordon et al. 2017; C. Zhang et al. 2017; Durr et al. 2018; R. Wu et al. 2018) and nucleases with numerous redundant targets

have been used to reduce the number of genes in a large family, such as the α-gliadins to create low-gluten, non-transgenic wheat (Sánchez-León et al. 2018). Use of tissue-specific promoters to express nucleases can be used to inactive genes in somatic cells in a tissue-specific manner, facilitating the analysis of gene function without the need to create a heritable event (Decaestecker et al. 2019).

Gene Targeting

Precise, homology-dependent modifications are typically referred to as gene targeting (Paszkowski et al. 1988). Gene targeting utilizes donor DNA - a sequence of DNA encoding desired genome alterations and dozens to hundreds of homologous flanking nucleotides - to create site-specific modifications to a DNA sequence. After DSB, resection reveals ssDNA homologous to the donor molecule. Homology-seeking proteins are loaded onto this ssDNA that bind and copy similar DNA sequences, resulting in the incorporation of the supplied donor DNA sequence into the genome. The delivery of a donor template paired with an SSN can result in that information being perfectly copied into the genome at a precise location (Figure 1-1E) (Lisby and Rothstein 2015). In plants, two homologous recombination (HR) sub-pathways (synthesis-dependent strand annealing and HR utilizing a Holliday junction intermediate) appear responsible for gene targeting (Holger Puchta 1998). This has implications for donor design, which has been thoroughly discussed elsewhere (Huang 2019). Studies outside of plants have indicated that HR is primarily used to incorporate information from double-stranded donors (Kan et al. 2014). Information from single-stranded donors (or the ssDNA itself in some cases) is likely incorporated by other pathways such as single-strand assimilation (SSA), often at much greater efficiency (Kan and Hendrickson 2019).

The greatest strength of HR-based approaches is the ability to incorporate any novel DNA sequence into the genome. However, HR occurs infrequently in somatic plant cells and NHEJ and MMEJ typically predominate. Most published approaches typically still require phenotypic markers to recover gene targeting

events due to low efficiencies and high variability of outcomes (Huang 2019). For example, many published gene targeting experiments involved creating herbicide tolerance, because it is possible to easily select herbicide tolerant cells that have undergone HR (Ayar et al. 2013; Schiml, Fauser, and Puchta 2014; Endo, Mikami, and Toki 2016; Kumar et al. 2016; Sun et al. 2016).

Multiple strategies have been pursued to increase gene targeting frequencies in plants. In several dicots (*Nicotiana benthamiana*, tomato and potato) and rice, 'replicons' consisting of replication elements from Bean Yellow Dwarf Virus (BeYDV) or Wheat Dwarf Virus (WDV), respectively, have been used to increase GT frequencies and create whole, modified plants (Baltes et al. 2014; Tomáš Čermák et al. 2015; Butler et al. 2016; Wang et al. 2017). Replicons are an abundant source of donor DNA and nuclease within the cell. They also induce S-phase, which is conducive to HR repair, yielding higher gene targeting frequencies (Baltes et al. 2014). In some studies however, replicons increased gene targeting frequencies, but the modified cells did not give rise to whole, modified plants (Gil-Humanes et al. 2017). It is possible that replicons compromise cell division in some species or under certain growth conditions. That is, cells containing functional replicons undergo gene targeting at a high frequency but are unable to propagate. This may be overcome in the future by using alternative, attenuated, or transient replicons. Additionally, it is important to employ the correct experimental strategy when using replicons. The relative position of transcriptional units on a replicon vector are sensitive to the position of the viral elements; replicons have bi-directional promoters resulting in some published replicons likely repressing the expression of the neighboring gRNA via the formation of dsRNA (Hahn et al. 2018).

Further success at enhancing gene targeting has been made using the *in planta* targeting approach, in which the donor is excised from the genome by the same nuclease that cuts the target (Ayar et al. 2013; Schiml, Fauser, and Puchta 2014; Kumar et al. 2016; Sun et al. 2016; Fauser et al. 2012; Zhao et al. 2016). The nuclease and donor are typically delivered to cells on the same construct,

and gene editing can occur at any point after transformation. Increases in efficiency likely result because the released donor DNA can move throughout the nucleus to better repair the cleaved target site.

Whereas the strength of homology-directed repair is its precision, other tools can be used to create precise genome modifications. Most prominent among these are base editors, such as cytosine and adenosine deaminases (Komor et al. 2016; Gaudelli et al. 2017; Kang et al. 2018; C. Li et al. 2018; Zong et al. 2018). Targeted base deamination creates SNPs in a small window, allowing for the semi-random diversification of a target sequence to modify promoter elements, splice sites, and start codons. Much work is being done to increase the precision of base editors, however for applications in plants, there may be value in introducing random mutations in a short window.

Recovery of edited plants

Genome editing reagents are typically delivered to plant cells by one of two means: bacteria capable of directly delivering DNA to plants (such as *Agrobacterium tumefaciens*,) or biolistic bombardment of gold particles (Sanford 1990; Lacroix and Citovsky 2019). Susceptibility of cells to Agrobacterium is a complex trait, resulting in highly-variable delivery efficiencies between ecotypes and species (Lacroix and Citovsky 2019). The second approach, biolistic bombardment, makes it possible to delivery anything that can be bound to gold particles (usually DNA, RNA, protein, or some combination thereof); however, delivery is often inefficient. Both Agrobacterium and particle bombardment are typically used to deliver reagents to somatic leaves or cotyledons (albeit with some notable exceptions including ‘floral dipping’ of Arabidopsis). While both of these approaches are adequate for transient assays (genome editing or otherwise), near-exclusive delivery to non-germinal tissues results in additional steps being needed to generate whole, germinally-modified plants from the somatic tissue (Altpeter et al. 2016).

Plants distinguish themselves from most complex eukaryotes in the totipotency of their tissues (Vasil and Vasil 1972). This has long allowed

researchers to convert sectors of somatic tissue into whole plants. This somatic-germinal conversion (or regeneration) is the foundation of most plant transformation approaches: transgenes are delivered to isolated somatic tissue (be it protoplasts, immature embryos, leaves) followed by selection for the transgene and regeneration of the modified tissue into a whole, transgenic plant (Figure 1-2A shows traditional *Agrobacterium*-mediated maize transformation as an example) (Cody, Graham, and Birchler 2017). Despite many of these protocols being developed over decades, the process is far from routine in most laboratories. Further, success is often genotype dependent, and the regenerated plants can have changes to their genome and epigenome (Phillips, Kaeppler, and Olhoft 1994; Kaeppler, Kaeppler, and Rhee 2000).

An emerging alternative to traditional regeneration techniques is somatic reprogramming, in which rather than inducing roots and shoots from callus, cell fate is modified by expressing morphogenic regulators (Figure 1-2B) (Lowe et al. 2016; Mookkan et al. 2017; Lowe et al. 2018). Somatic embryogenesis, or the generation of embryos from somatic tissue, has been performed by the expression of two developmental genes, *BABYBOOM* (*BBM*) and *WUSCHEL* (*WUS*); a developmental factor long associated with somatic-embryogenesis and a well-studied meristem maintenance transcription factor, respectively (Lowe et al. 2016; Mookkan et al. 2017; Lowe et al. 2018). Since being initially performed and published in maize, the use of *BBM* and *WUS* for somatic embryogenesis has gone through several phases of development and has been shown to be extremely efficient and widely-applicable in the hands of its developers.

Another approach utilizing the delivery of morphogenic or developmental regulators induces meristem formation (Figure 1-2C) (Maher et al. 2020). For example, when *WUS* and a cytokinin biosynthesis gene are co-delivered with gene editing reagents to dicots, new meristems are created. Some of the meristems have gene edits, allowing for both tissue culture-free and transgene-free gene editing. The advantages of morphogen-based approaches over conventional regeneration techniques are numerous, most notably the speed

(fewer treatments/media changes), number of plants recovered, reduction in somaclonal variation, and broad applicability. Another advantage is that the formation of embryos or meristems itself serves as a transformation marker – making escapes virtually impossible and thereby significantly reducing the effort required during the screening process. It is clear that these types of approaches that minimize or eliminate tissue culture are needed to open the current bottleneck in plant gene editing. Development, dissemination and adaptation of such protocols should be a high priority for the gene editing community.

Conclusions

The advent of genome editing is revolutionizing the life sciences and greatly advancing basic and applied plant biology. However, bottlenecks need to be overcome before the full potential of this technology is realized in plants. The development of improved gene targeting strategies, replicons, base editors and targeted non-homologous insertions all show great promise for eliminating the bottleneck for making precise gene edits. The use of developmental regulators promises to greatly increase the pace at which modified plants may be created, perhaps obviating the need for tissue culture in the near future. Together, these advances and others will ensure the real potential of gene editing in plant biology is fully realized.

Acknowledgements

We thank M. Leffler for help with the figures.

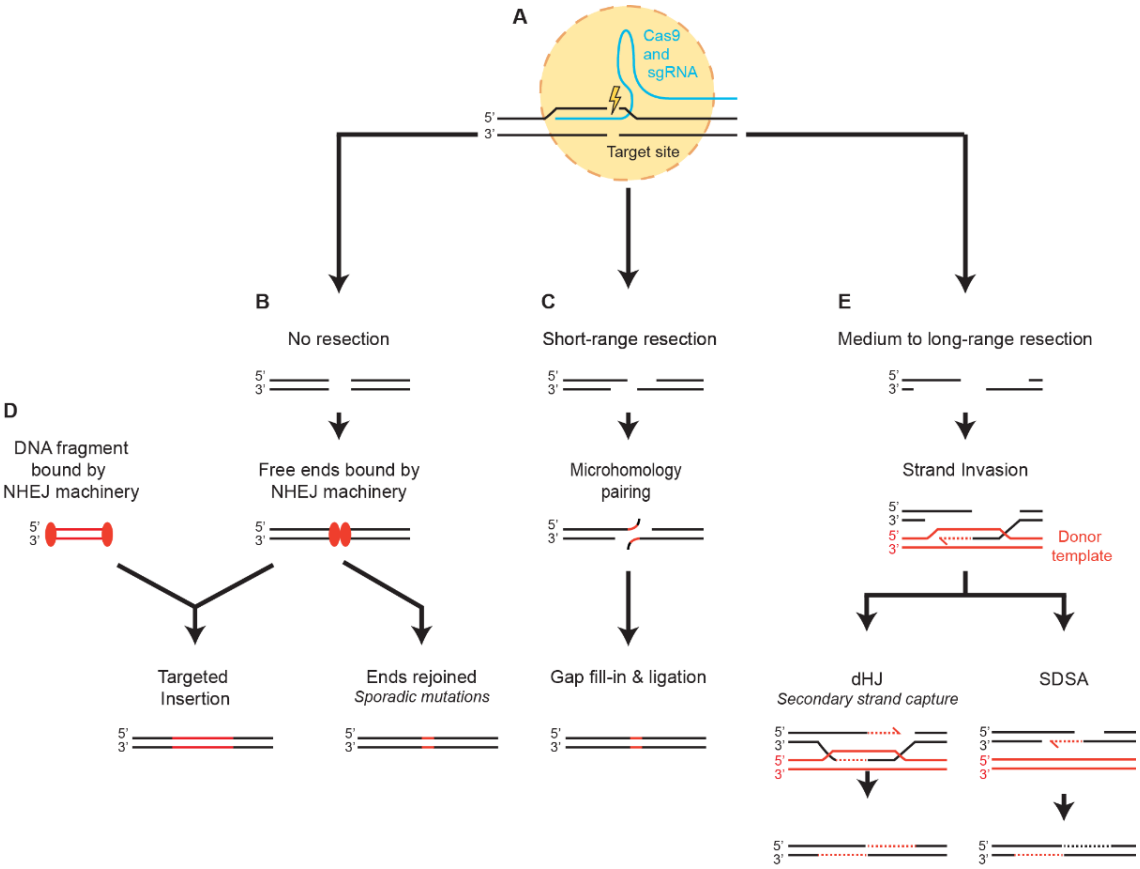

A
Cas9
and
sgRNA
5'
3'
Target site
B
No resection
C
Short-range resection
E
Medium to long-range resection
D
DNA fragment
bound by
NHEJ machinery
Free ends bound by
NHEJ machinery
Microhomology
pairing
Strand Invasion
Donor
template
Targeted
Insertion
Ends rejoined
Sporadic mutations
Gap fill-in & ligation
dHJ
Secondary strand capture
SDSA

Figure 1-1 DNA damage-mediated genome editing is facilitated by a variety of repair outcomes. (A) Cas9 binds and creates a DSB at a genomic target site. (B) Free DNA blunt ends are bound by non-homologous end joining (NHEJ) machinery. These ends are then rejoined, typically in an error-free manner or in some cases joined to another DNA fragment bound by NHEJ-machinery. (C) In microhomology-mediated end joining (MMEJ), end-processing machinery resects 5' ends a short distance exposing 3' ends that anneal with minimal homology (red region). Error-prone polymerase and DNA processing enzymes complete the repair, frequently resulting in small insertions and deletions. (D) Linear DNA fragments may be inserted into DSB sites, allowing for targeted insertions without homology. (E) Homologous recombination edits are initiated by 5' resection followed by a homology search with the free 3' end. Upon binding to a homologous sequence, the 3' end is extended, copying the donor template (red dashed line). Secondary-strand capture results in the formation of a double Holliday junction (dHJ) which is then resolved. Alternatively if the synthesis-dependent strand annealing (SDSA) pathway is utilized, the invading strand dissociates from the donor template and incorporates into the target site. dHJ and SDSA products differ in that dHJ can copy donor information in both directions relative to the target site (red dashed lines) while SDSA is unidirectional.

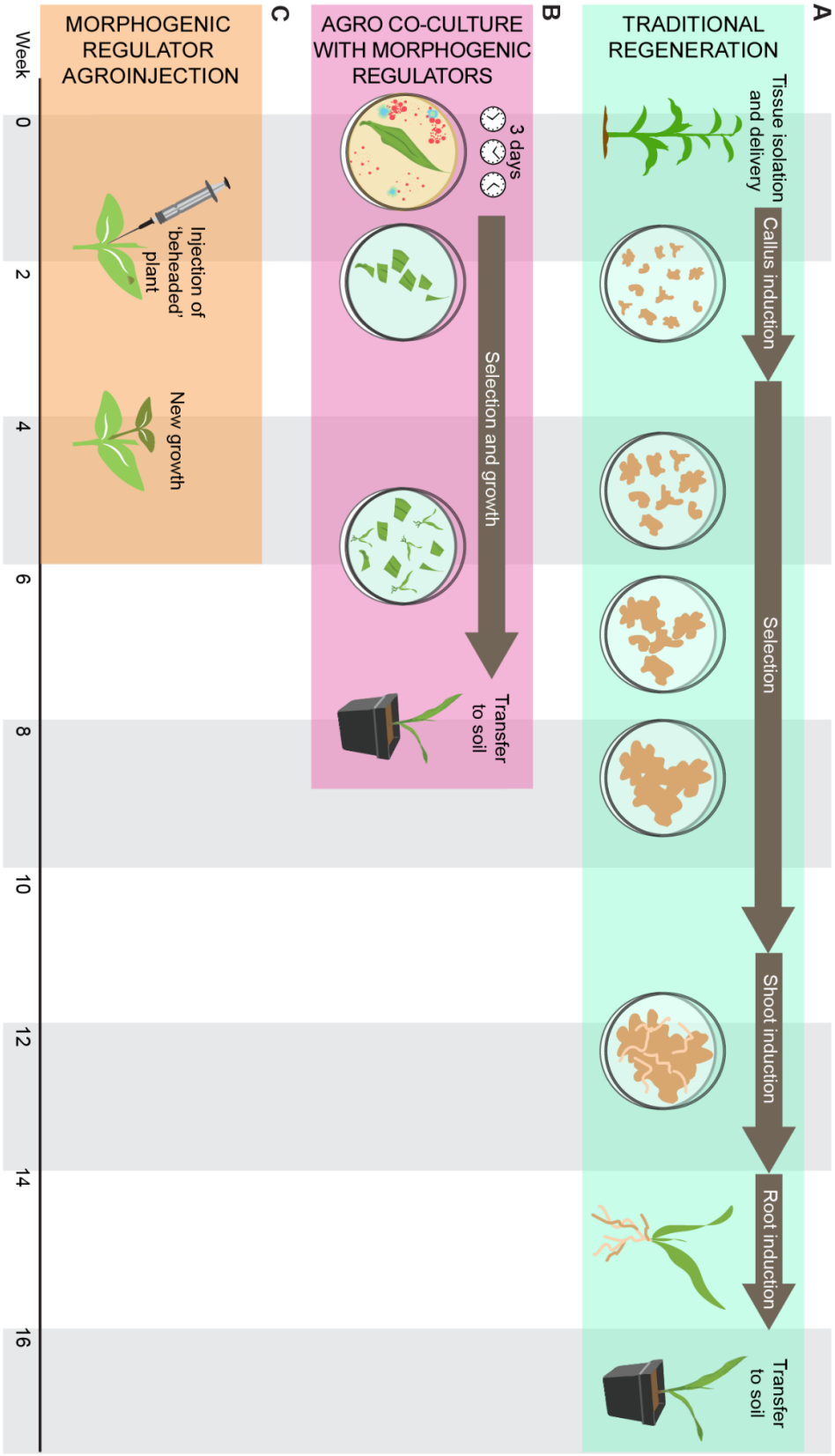

A
TRADITIONAL REGENERATION
Tissue isolation and delivery
Callus induction
Selection
Shoot induction
Root induction
Transfer to soil
B
AGRO CO-CULTURE WITH MORPHOGENIC REGULATORS
3 days
Selection and growth
Transfer to soil
C
MORPHOGENIC REGULATOR AGROINJECTION
Injection of 'beheaded' plant
New growth
Week
0
2
4
6
8
10
12
14
16

Figure 1-2 Timeline comparison of emerging and traditional transformation techniques. Time required to generate starting plant material and seed-setting after generation of new material not included to highlight differences in time to recover modified plants. (A) Traditional *Agrobacterium*-mediated maize embryo transformation. Maize embryos are extracted and incubated with agrobacterium. Multiple stages of selection are used to identify transgenic callus. Selected callus are transferred to shoot induction media follow by root induction media. After sufficient root growth, plants are transferred to soil. (B) Maize transformation using emerging morphogenic tools. Maize Leaves are cut and co-cultured with *Agrobacterium* for 3 days. Morphogens induce the formation of embryos that develop over 7-8 weeks after which they are transferred to soil. (C) *Nicotiana benthamiana* transformation via morphogenic regulator agroinjection. Plants are 'beheaded' to remove apical dominance and the stem is agroinjected. New growths that appear after 3-5 weeks are strong candidates for having been delivered the morphogenic T-DNA cargo.

CHAPTER TWO

Methods for Analysis of Somatic Plant Gene Targeting Events

Paul A.P. Atkins, Xu Tang, Redeat Tibebu, and Daniel F. Voytas.

Preface

Numerous assays have been used to measure gene targeting frequencies in plants, including molecular and phenotypic readouts from genomic, synthetic, integrated, and episomal targets. This wide variety in assays has resulted in studies producing relative results that are difficult to translate between platforms, even within a plant species. A priority of the plant genome editing community must be the standardization of assays to allow for cross-platform comparisons and to enable largescale optimization. Here we evaluate methods for analyzing gene targeting outcomes in search of a high-throughput, scalable and quantitative assay that could be broadly adopted by the plant genome editing community.

Introduction

Optimization of plant gene targeting has progressed slowly compared to other biological systems (Huang and Puchta 2019; Altpeter et al. 2016). Whole-plant transformation methods are restricted to low per-cell approaches (biolistics) or low copy number approaches (Agrobacterium), and regeneration is technically challenging for all but the most specialized laboratories. While biolistics and Agrobacterium are limited, one or both delivery methods are routine for research groups engineering plant genomes. This, combined with the ease of creating CRISPR-based genome editing reagents, has resulted in the delivery of gene targeting reagents to somatic plant tissue a relatively straightforward task for those in the field (T. Čermák et al. 2017; Harwood, Volpi e Silva, and Patron 2017; Hahn et al. 2019). Today, what is lacking is a high-throughput platform for analyzing gene targeting events in somatic cells to facilitate optimization and mechanistic dissection. Here we examine current methods for studying plant gene targeting and their potential as in a high-throughput, broadly applicable platform.

Key Properties of a Gene Targeting Analysis Platform

A high-throughput and robust platform for assessing plant gene targeting is needed. In plants, dozens or hundreds of somatic assays (e.g. delivery to leaf tissue) can be performed in 2-5 days. This is vastly faster than ‘whole plant’ approaches in which a germinally-modified plant is produced, requiring multiple generations and technical proficiency (Altpeter et al. 2016; Simmons, VanderGheynst, and Upadhyaya 2009; H.-Y. Wu et al. 2014). Somatic cell experiments may favor approaches that do not function in germinal/regenerating tissue, but the enormous throughput can be harnessed to reduce treatments tested in a ‘whole plant’ system. An ideal somatic assay produces data that is readily quantifiable and has a linear dynamic range to capable of capturing low-frequency events. This can be accomplished by counting individual cell-autonomous events (e.g. localized fluorescent protein, sequencing of target site) or non-cell autonomous events, such as measuring the production of a gene targeting-dependent product (e.g. quantitative measurements of enzyme activity

such as β-glucoronidase (GUS) from leaf punches). In either case, the readout may be the direct consequence (tagging endogenous gene with GFP or sequencing a modified DNA sequence) or indirect consequence (modification of regulatory sequence that results in GFP expression) of gene targeting. Existing tools for detecting gene targeting can be categorized into phenotypic and molecular methods.

Phenotypic Reporters: Cell, Tissue, and Developmental Phenotypes

Phenotypic methods rely on the editing event to create a readily-detectable change in the plant, such as the precise insertion of a GFP cassette into a highly expressed gene or the modification of an herbicide tolerance gene (e.g. acetolactate synthase, *ALS* and plant polyphenol oxidases, *PPO*) to create a resistance phenotype (Beetham et al. 1999; Hanin et al. 2001; Shaked, Melamed-Bessudo, and Levy 2005). Phenotypic methods can be divided into those that confer cell-autonomous phenotypes, tissue level phenotypes, or whole-plant phenotypes, each of which requires distinct experimental approaches. A cell-autonomous reporter, or a reporter that can be detected in an individual cell following the gene targeting event, can be assayed in any tissue regardless of the mutation's presence during development. These reporters are often visual, e.g. fluorescent (GFP), bioluminescent (luciferase), or enzymatic (GUS). A tissue level phenotype is one that can be readily detected after delivery of reagents, but the measurement cannot be directly tied to an individual cell that underwent gene targeting; e.g. the measurement of a metabolic product in a cell lysate. Differences between cell autonomous and tissue level phenotypes is often the detection method rather than the gene product itself (GUS measurements with microscopy of intact tissue compared to a plate reader and lysate). A whole-plant phenotype requires the modification to be present during development and/or be highly prevalent (far above expected gene targeting frequencies) to detect a phenotype, which include herbicide tolerance and developmental alterations (e.g. plant architecture, leaf shape, trichome formation). This requirement is often incompatible with transient delivery methods and may

require the regeneration of uniformly modified tissues, making it drastically lower throughput.

Phenotypic Reporters: Minimizing Type I Error

A phenotypic gene targeting reporter must not be prone to spurious activation; the false positive rate must be extremely low due to the expected low frequency of plant gene targeting in some systems (below 0.1%). This imposes restrictions on donor design whenever the reporter's full coding sequence is contained within the donor -- the donor should not contain promoter elements, start codons, or translation initiation sites that may result in a functional gene product. Targeting of exons beyond the first reduces the likelihood of including minimal or cryptic promoter elements into the donor arm (e.g. 5' donor arm positioned to contain an intron rather than promoter sequence), with 3' translational tagging perhaps being the most conservative approach. Targeting of interior exons is not ideal due to the possibility of single-sided gene targeting events (synthesis-dependent strand annealing [SDSA] insertion) in which either the 5' junction contains mutations shifting the reporter out of frame or where the 3' junction contains a mutation that shifts downstream exons out-of-frame. Both cases may trigger nonsense-mediated decay of the transcript. Single-sided SDSA events can be accommodated by targeting an entire synthetic exon into an intron or by including a terminator sequence at the 3' of the reporter sequence, or a combination of both. The use of a synthetic exon requires several genomic context-dependent controls to ensure its function and the frequency of background, and inclusion of a terminator increases the size of the targeted insertion. Together, these conditions result in a bias towards targeting high-expression, multi-exon genes with low tissue-specific expression and requires delivery of whole reporter gene sequences.

Phenotypic Reporters: Targeting Restrictions

Phenotypic gene targeting reporters further restrict target sites by requiring a dominant phenotype. Target include herbicide resistance genes (e.g. ALS and PPO), non-functional endogenous genes (of previously mutated by

genome editing, e.g. restoring trichomes), pigment genes (e.g. overexpression of anthocyanin), and highly expressed loci (e.g. *GFP-CRUCIFERIN* or *GFP-ACTIN* translational fusion) (Paszkowski et al. 1988; Beetham et al. 1999; Hanin et al. 2001; Shaked, Melamed-Bessudo, and Levy 2005; Saika et al. 2011; Tomáš Čermák et al. 2015; Hahn et al. 2018). A target's basic gene function must be understood and is most likely to be a protein-coding gene.

An alternative to endogenous targets are integrated reporters: transgenes engineered to give a specific outcome when modified by genome editing tools. An integrated reporter construct ensures consistent expression, and minimizes the modification required for activation (e.g. encodes 90%+ of a reporter protein sequence). Most integrated reporters have been selected for a chromatin context amenable to high gene expression during transformation/regeneration and therefore cannot be used to assess other chromatin contexts unless exceptional effort is made to recover non-selected, random integration events (Kim and Gelvin 2007).

A target site may also be co-delivered with the gene targeting reagents, typically seen in protoplast-based or biolistic experiments where many copies of plasmids can be expected to be delivered to each cell (Endo, Mikami, and Toki 2016; Terada et al. 2002; Shan et al. 2014). Interpretation of these experiments may be confounded by the differences in copy number and chromatin context between a delivered DNA fragment and an endogenous target. Use of endogenous, pre-integrated or co-delivered phenotypic reporters results in biases in target site that makes them undesirable for use in an optimization pipeline, which should function at any conceivable target.

Molecular Detection: A Note on PCR Bias

Molecular detection methods do not require the expression of a gene product but still impact the design of gene targeting reagents. Virtually all high-throughput molecular approaches require an amplification step: PCR. This introduces bias into all downstream analysis (M. T. Suzuki and Giovannoni 1996; Polz and Cavanaugh 1998; Acinas et al. 2005; Pinto and Raskin 2012; Eloe-

Fadrosh et al. 2016; Silverman et al. 2019). Perhaps the most relevant is size bias – any amplicon population that includes significant size variation will see the smaller products more efficiently amplified and overrepresented in the final amplicon pool. This must be considered when designing reagents, as very small gene targeting insertions mimicking protein tags or a small functional domain would not be expected to significantly impact analysis while large insertions may. Large insertions do not necessarily prevent quantification of events but require additional controls to determine how input and output ratios differ for those products (Kalle, Gulevich, and Rensing 2013).

Beyond differences in amplification efficiencies due to size, PCR is biased by its uneven sampling of any population, even in the case of identical amplicons. This uneven sampling is due to random variation that occurs during PCR cycling – each individual starting molecule and its descendants are stochastically amplified during each cycle and these minute differences compound over many cycles, resulting in some sequences being significantly overrepresented in the population, or jackpotting. These jackpotting events do not prevent analysis (PCR is routinely used for quantification), but they most noticeably impact rare or unique events in a population; any small variation in amplification efficiency for a unique event can drastically alter its prevalence in the population while a well-represented sequence may have identical sequences both over and underrepresented. Jackpotting events may be minimized when searching for rare events by pooling redundant amplifications - the stochastic nature of jackpotting will result in different sequences being favored in each amplification and the pool may better represent the starting population, or at least sample a greater diversity of events. Additionally, unique molecular identifiers (UMIs) can be used to tag each starting template to allow for sequences derived from the same starting template to be combined (Kivioja et al. 2012; Karst et al. 2020).

Primer Placement

Reliance on PCR requires careful primer placement to minimize false-positives – any region homologous to sequences within the donor molecule should be avoided to prevent priming. Even if a functional amplicon cannot be formed from a primer binding site within the donor (e.g. both primers binding the Watson strand of the target), this should be avoided due to the creation of linear fragments that can recombine via template switching during PCR and create amplicons that are indistinguishable from the expected gene targeting event. Experimentally determining the probability of template switching events is technically challenging, making it better to try to reduce the potential for these events and perform a template switching control (Potapov and Ong 2017). A donor containing SNPs without a nuclease can be used to detect template switching events; SNPs deposited in patterns not consistent with known gene targeting mechanisms (e.g. distal SNPs being incorporated preferentially) or above expected GT frequencies are putative PCR artifacts.

Gene Targeting-Specific PCR

One common approach for molecularly detecting gene targeting events is the gene targeting-specific PCR (Koller and Smithies 1992). In this, one primer is targeted to the donor insertion/sequence modification and the second binds endogenous sequence outside the homologous region. Ideally, the PCR can only produce an amplicon if the gene targeting modification has been incorporated into the target site and, due to PCR's extremely low limit of detection, it may be possible to selectively sequence a lone gene targeting event in a population of WT sequences. In practice, residual donor molecules contaminating the genomic DNA template may confound the result by creating false positives via template switching. Template switching occurs when each oligo creates a linear product that possess sufficient complimentary sequence to bind in subsequent PCR cycles; after 2 cycles of PCR it is possible for multiple copies of a perfect gene targeting event to be present in the PCR despite it being absent in the starting material. Gene targeting PCR's propensity for false-positives when the donor is present (as is the case in any transient plant system) and its non-quantitative

nature (only presence/absence) makes it only useful as a tool for preliminary analysis of gene targeting experiments or when the no donor is present (e.g. screening regenerated lines).

PCR Digest

A simple method for detecting gene targeting is the PCR digest, in which the target is PCR amplified and incubated with a restriction enzyme that is specific to either the unmodified target or the gene targeting modification. With the former, any amplicons cut by the restriction enzyme contain an intact target site and the uncut amplicon has been modified by the treatment. With the latter, only amplicons derived from the expected gene targeting event are cleaved. A combination of both yields both gene targeting and targeted mutagenesis frequencies. While this is often viewed as a cursory analysis method, it is only limited by the DNA fragment visualization technique used and restriction enzyme efficiency. A typical agarose gel will result in a poor limit of detection and imprecise quantification, but an Agilent Bioanalyzer is capable of precise quantification. Further, use of custom CRISPR-based restriction enzymes *in vitro* means this approach may no longer be limited by the presence of a 'standard' restriction enzyme site at the target. The only significant limitation of this method in the context of an optimization platform is depth – only presence/absence of a specific sequence can be detected, precluding any analysis that seeks to characterize all outcomes at a target site.

qPCR and Probe-Based Approaches

qPCR and qPCR-derived techniques (probe and dye-based qPCR, digital droplet PCR) all require optimized PCR conditions and are restricted to short amplicons and therefore incompatible with large donor molecules (Day, Dear, and McCaughan 2013). Shortening homology arms to accommodate smaller amplicons may facilitate analysis, but large donor arms are the standard in plant gene targeting experiments and appear to be required in many contexts. In contexts where donor arm length is not an issue (e.g. small oligo donors delivered to protoplasts or via biolistics), these approaches are appropriate, but

size restrictions prevent their use in a pipeline aspiring to compare all possible reagent variants.

Sanger Sequencing

Two distinct Sanger sequencing-based methods are frequently used to detect gene targeting events. First is the sequencing of amplicons that have been cloned into a vector. This requires PCR amplification of the target site, cloning of the amplicon into a vector of choice (often 'TA' cloning'), and Sanger sequencing of those plasmids. While this is quantitative and can yield a large amount of information for each target site, it is extremely low throughput; each read requires a transformed colony and generates its own Sanger trace. Scaling this approach is possible today, but it is extremely low throughput, expensive, and its advantages are largely shared by 3rd generation sequencing. Efforts to enrich the amplicon population for gene targeting events (e.g. digestion of WT sequences) render the assay non-quantitative and therefore not useful in an optimization pipeline.

TIDE (Tracking Indels by Decomposition), ICE (Inference of CRISPR Edits), and EditR (Edit deconvolution by inference of traces in R), are analyses performed on Sanger traces collected from a population of cells delivered genome editing reagents (Kluesner et al. 2018; Brinkman and van Steensel 2019). These tools allow for estimates of targeted mutagenesis and gene targeting by tracking the 'decomposition' of the trace that occurs at the target site. These tools have seen widespread use but their poor limit of detection and the lack of depth (no sequences from any individual events) make it another tool for cursory analysis.

Illumina Sequencing

Illumina sequencing readily generates massive amounts of data. Genome editing outcomes from millions of cells can be analyzed with ease by sequencing an amplicon of the target site. Its shortcomings become apparent when designing amplicon primers to assess gene targeting outcomes – Illumina reads are significantly shorter than standard donor arms used in plants. This results in few

to no possible primer sites that will result in sequence from the target site, with reads either being unable to reach the target site (sequencing only the homology arms) or primers being placed within donor molecules resulting in false positives (Figure 2-1A-B).

Bypassing Size Limitations

Illumina sequencing's core technical issue is its initiation from the ends of amplicons at primer binding sites added during PCR or via ligation. This results in reads beginning outside the donor arms and not spanning target site. A Unique Molecular Identifier (UMI) sample preparation and a modified UMI amplicon preparation protocol may allow for the placement of the sequencing initiation sites within the donor arm, allowing short reads to sequence the target site regardless of donor arm size. UMI protocols utilize a PCR with minimal cycles using only a single primer, or linear extension, to individually tag starting template molecules with a DNA randomer within each primer (5' 10-14bp random sequence) (Figure 2-2A). During thermocycling, the single oligo will bind and extend, creating a single-stranded product originating from its binding site (Figure 2-2B). Each additional cycle will create copies from the starting template, but because the single stranded product cannot serve as template for itself, the amplification is not exponential. The linear products are purified and further amplified with the standard, two oligo approach, wherein 1 oligo binds to a sequence tag incorporated in the linear extension oligo and the second oligo binds within the linear extension product (Figure 2-2C). This results in exponential amplification and the UMNIs allow reads with identical or highly-similar randomer sequences to be collapsed into a single sequence, minimizing bias produced by PCR cycling. The UMI sample preparation allows for one primer in the exponential amplification stage to be placed within the donor arm, but only if the donor molecule can be completely removed during the purification of the linear extension product. Additionally, the linear extension step makes each analysis strand-specific, potentially clarifying genome editing outcomes that

may be obfuscated by simultaneous amplification of both strands in a standard PCR amplification.

While the UMI preparation allows for the placement of one Illumina sequencing initiation site within the donor arm, this results in a small analysis window not centered on the target site (Figure 2-2D). This may be remedied by adding an additional step that transposes the UMI from the original primer used for linear extension to the oligo used in the exponential amplification step via circularization (Figure 2-3A-B). The subsequent exponential amplification is circularization-specific, with both oligos sitting within the donor arms, allowing for the entirety of the paired-end Illumina read to be centered on the target site (Figure 3C-D).

In practice, both the UMI and the UMI-circularization sample preparations created levels of technical artifacts that made rare event detection infeasible. Linear extension conditions were tested at two loci (one *Oryza sativa* endogenous sequence, *PDS*, and one transgene, *GU:PTII*) from untreated genomic DNA samples (Figure 2-4). The linear extension was found to require a large amount of DNA when using 20 amplification cycles. Next, these protocols were tested on *Oryza sativa* (rice) protoplast samples delivered nucleases and donor molecules targeting *PDS*. The delivered donor molecules contained regularly spaced SNPs to enable mechanistic analysis and to identify PCR artifacts. Illumina sequencing of the final amplicons derived from either the linear extension product or the circularized linear extension product revealed significant issues; the linear extension product was unable to be separated from the delivered donor molecule using standard cleanup methods resulting in template switching (Figure 2-5). This SNP deposition pattern can be attributed to template switching because SNP incorporation frequency decreases linearly from the primer binding site within the donor. The decrease is linear because the polymerase is equally likely to ‘fall off’ the template at any point and only these amplicons from which the polymerase has ‘fallen off’ can serve as primers on the linear extension product. This was further highlighted by the similar gene

targeting frequencies found in both the nuclease and non-nuclease treatment, a result known to not be biologically possible in this system (Figure 2-5). The data shown is the most severe example found using the UMI-circularization method, but the trend was also found at much lower (single digit) yet still debilitating frequencies in the UMI-only preparation (data not shown). Steps were taken in attempt to make the linear extension purification step more specific and to eliminate the donor molecule, but they were not found to be effective, with options being limited due to the need to preserve the ssDNA linear extension product. These included the use of magnetic beads (non-selective) and biotinylation/streptavidin purification of the linear extension oligo. Ultimately, these attempts to bypass the size limitations of Illumina sequencing were found to be technically challenging and not robust.

Third Generation Sequencing Techniques

Third generation sequencing techniques, Oxford Nanopore and PacBio SMRT (single molecule real-time) sequencing, obviate primer placement restrictions of Illumina by not limiting read size. Instead they are restricted by their read quality and throughput, respectively. Oxford Nanopore reads have been steadily increasing in quality for over a decade, but routine applications still struggle to exceed 92% accuracy, and the error is exacerbated by homopolymers (Laver et al. 2015; Jain et al. 2018). This necessitates analyses not dependent upon precise alignments or detection of gene targeting by a straightforward search for the sequence of interest. PacBio SMRT sequencing can generate high-quality, long sequences but is hamstrung by its per-read cost and variable read quality that decreases with amplicon size. This technique has been used to analyze gene targeting outcomes, and perhaps in the near future, wider adoption will decrease costs and increase accessibility such that it becomes a standard method for assessment of gene targeting outcomes (Hendel et al. 2014).

Conclusions

Given the strict parameters of a plant gene targeting optimization platform and issues with detection methods outlined here, third generation sequencing

techniques are the most promising option for developing a broadly-applicable platform for dissection and optimization of gene targeting in plants. Current off-the-shelf tools (both sequencing and bioinformatics) are sufficient for assessing genome editing outcomes with PacBio SMRT sequencing, but experiments are costly and low throughput, placing them out of reach of many labs. The significantly lower per-sequence cost and trivial overhead of Oxford Nanopore makes it ideal for the plant research community, but the lack of bioinformatic tools to parse desirable outcomes from the platform's significant noise hinder its application. This has created an opportunity for the development of a bioinformatic pipeline, allowing for the rapid assessment of plant genome editing outcomes and optimization of gene targeting strategies using a broadly applicable platform.

Acknowledgements
I would like to thank members of the Voytas lab and the University of Minnesota Genomics Center for the numerous discussions that guided this work.

Author Contributions
PAPA and DFV conceived and planned the research. PAPA wrote the text and carried out the analysis and experiments. NG, XT, and RT contributed materials and performed protoplast experiments.

Methods
Vector Construction: *GUS* repair control vectors (pLSLZ.D) were constructed as previously described. An sgRNA ('R2') targeting the *gu:ptii* target region was cloned into pTC213 using oligos oPAA394 (GTACGCGTCCCGGGTCGCTACCTT) and oPAA395 (AACAAGGTAGCGACCCGGGACGC) create pPA205 as previously described. Donor molecules of varying arm length (1kb, 250bp, 50bp) were PCR amplified from pLSLZ.D Cas9 and cloned via Gibson assembly into pTC214 after gel purification of the BaeI-digested backbone resulting in pPA206, pPA209, and pPA210 respectively. These donor plasmids were then each combined with pMODA0101, pPA205, and pTRANS_101 via Golden Gate assembly (Tomas

2017). This resulted in pPA216, pPA219, and pPA220, which all contained 35s:Cas9, AtU6:R2, and one donor molecule (1kb, 250bp, and 50bp donor arms, respectively).

Plant Material: *Nicotiana tabacum* var *Xanthi* encoding a broken *GUS* reporter transgene were grown at 22 celcius, in 16 hour/8 hour light/dark cycle (Wright et al. 2005). True leaves were used for leaf infiltrations, each leaf being infiltrated with the experimental treatment on one half and the control plasmid (pLSLZ.D) (Baltes 2015) on the second half (infiltration zones separated by midrib vein).

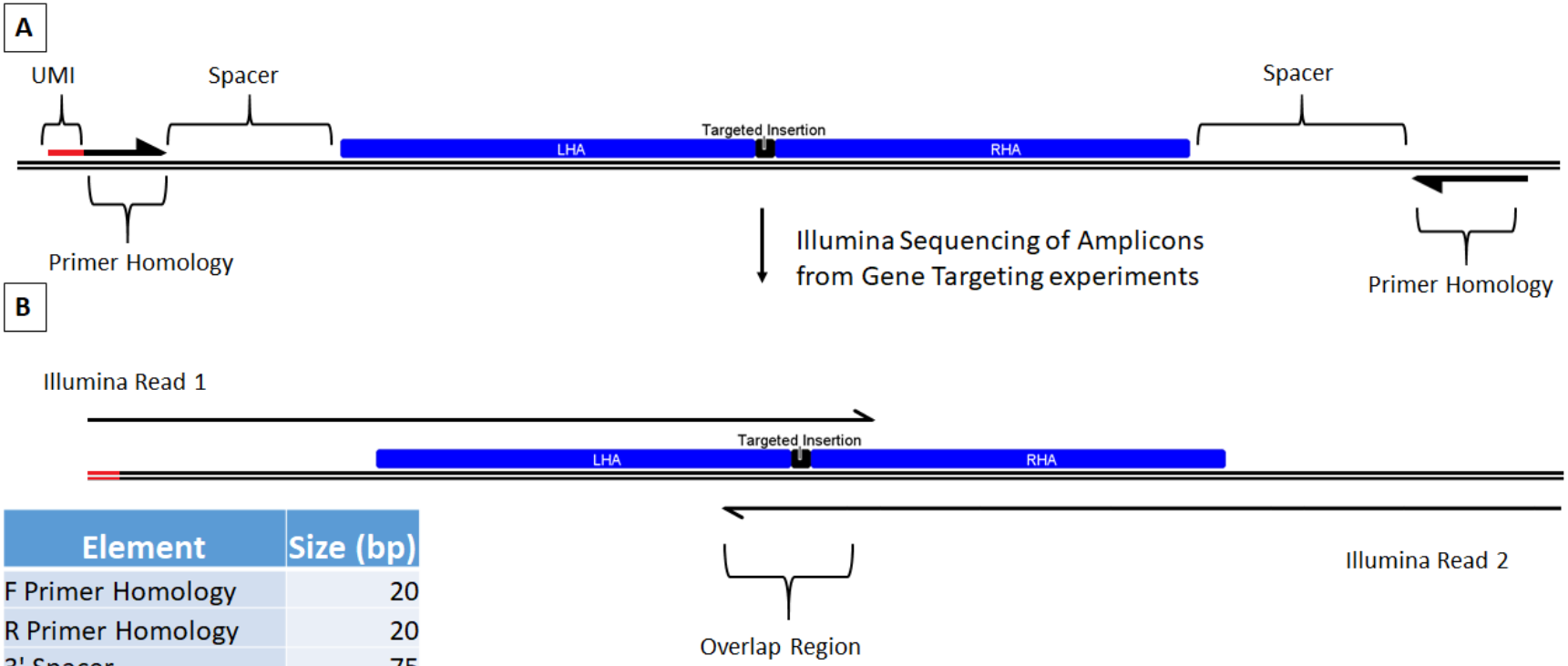
Nicotiana tabacum var *Xanthi* and *Oryzae sativa* protoplasts were grown, isolated, and PEG-transformed as previously described (Wright et al. 2005; Shan et al. 2013).

UMI Preparations: Linear extensions were performed using KapaHiFi HotStart polymerase. Genomic DNA input concentrations were measured using a Nanodrop. Cycling conditions followed manufacturer specifications with the exception of linear extension, which all utilized 20 cycles. Purifications performed on completed linear extension product included QIAquick PCR Purification Kit (catalog 28106 Qiagen), Agilent AMPure XP Beads (catalog A63880 Beckman Coulter), and, when using biotinylated oligos (Integrated DNA Technologies) streptavidin beads (catalog S1420S New England Biolabs).

Circularizations were performed using both a blunt-end approach and an adapter-mediated overhang ligation following restriction enzyme treatment. After purification a two cycle PCR was performed to create dsDNA for downstream circularization via ligation. In one case, oligos containing adapters with restriction enzyme sites that created sticky overhangs after incubation with a restriction enzyme after cycling. These products were again purified and treated with NEB Quick Ligase per the manufacturer's instructions. This ligation product was used as a PCR template using KapaHiFi Hot Start using circularization primers that contained Illumina sequencing adapters.

Illumina Sequencing: Illumina Sequencing was performed by GeneWiz using their AmpliconEZ service.

Bioinformatic Analysis: Illumina reads were analyzed using an early version of PANGEA described in Chapter 3.



Element	Size (bp)
F Primer Homology	20
R Primer Homology	20
3' Spacer	75
5' Spacer	75
UMI	14
F Overlap	30
R Overlap	30
Total Size	**254**

Figure 2-1 Limitations of paired-end Illumina sequencing when analyzing plant gene targeting experiments. (A) Schematic of amplicon elements. UMI is a unique molecular identifier, primer homology refers to the portion of the primer that binds a targeted sequence, and spacer refers to a region between primer binding sites and donor homology region to minimize template switching. LHA and RHA are left homology arm and right homology arm, respectively, designating the donor homology regions relative to the target site. Targeted insertion is the gene targeting cargo being copied into the target site via homologous recombination. (B) Illumina paired-end sequencing of gene targeting amplicons. Illumina read 1 and 2 initiate from the ends of the amplicon, sequencing towards the center. The table displays the approximate size of each of the amplicon elements that must be sequenced except the homology arm and insertion. These values can be used to estimate the upper size limit of insertions and donor arms that can still be accommodated. For example, 2x300 (600 starting bases) sequencing leaves 346 bases remaining for the homology arms and donor insertion, heavily restricting gene targeting reagents.

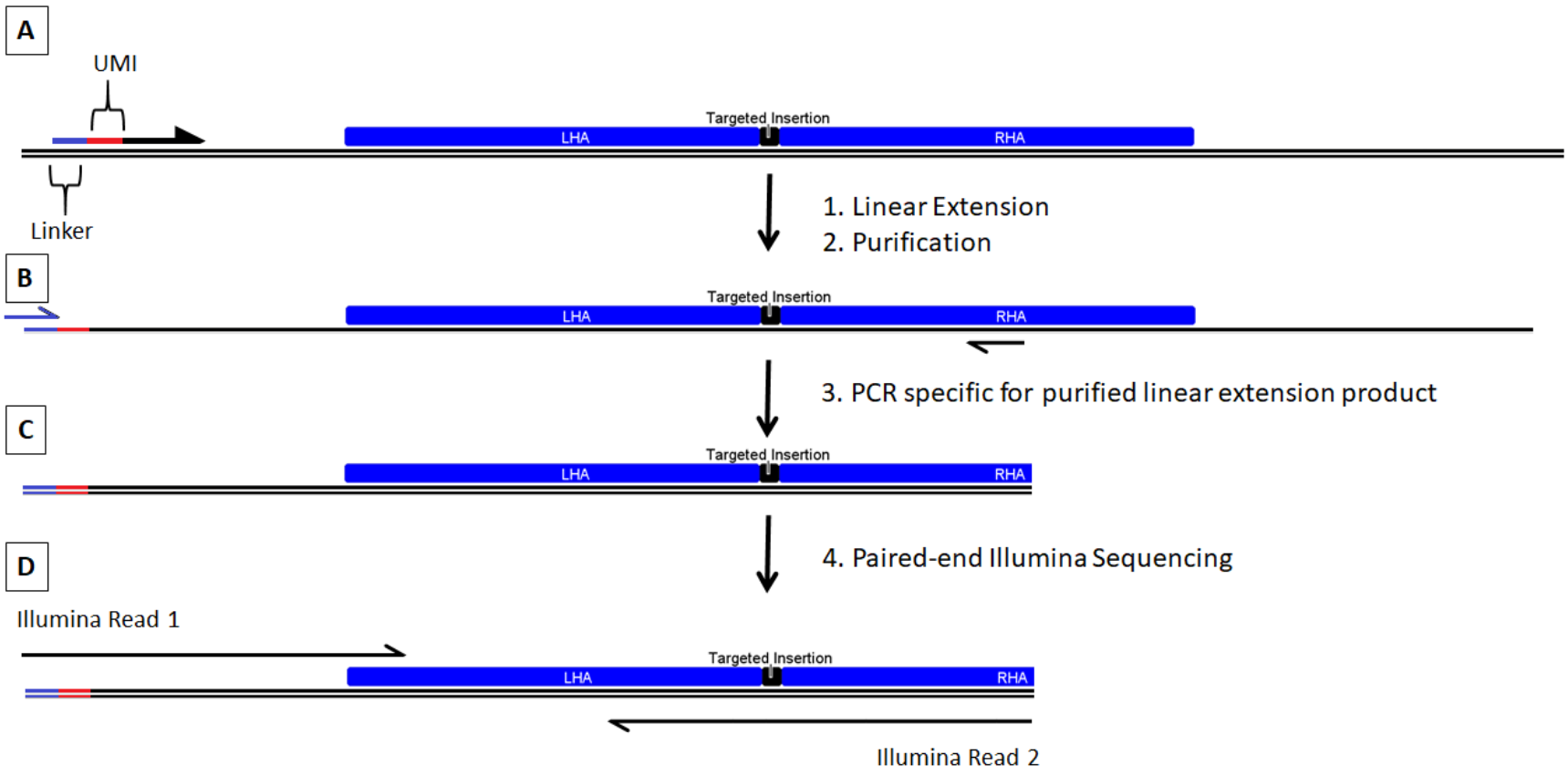

A
UMI
Linker
Targeted Insertion
LHA
RHA
1. Linear Extension
2. Purification
B
Targeted Insertion
LHA
RHA
3. PCR specific for purified linear extension product
C
Targeted Insertion
LHA
RHA
4. Paired-end Illumina Sequencing
D
Illumina Read 1
Targeted Insertion
LHA
RHA
Illumina Read 2

Figure 2-2 UMI sample preparation may bypass Illumina size restrictions. (A) Linear extension step. A single oligo containing a UMI (typically a 14bp randomer sequence) and a linker for downstream amplification is used to create ssDNA copies of the target region by thermocycling. (B) Purified linear extension products are selectively amplified using a primer specific to the initial linear extension oligo linker region. The second oligo is placed within the donor homology region ensuring the NGS read spans the target site and any expected modification. (C) Double-stranded amplicons are prepared for sequencing using standard Illumina pipelines. (D) Amplicon sequencing resulting from this strategy is deduplicated using UMI sequences, and the reads spanning the target site are analyzed.

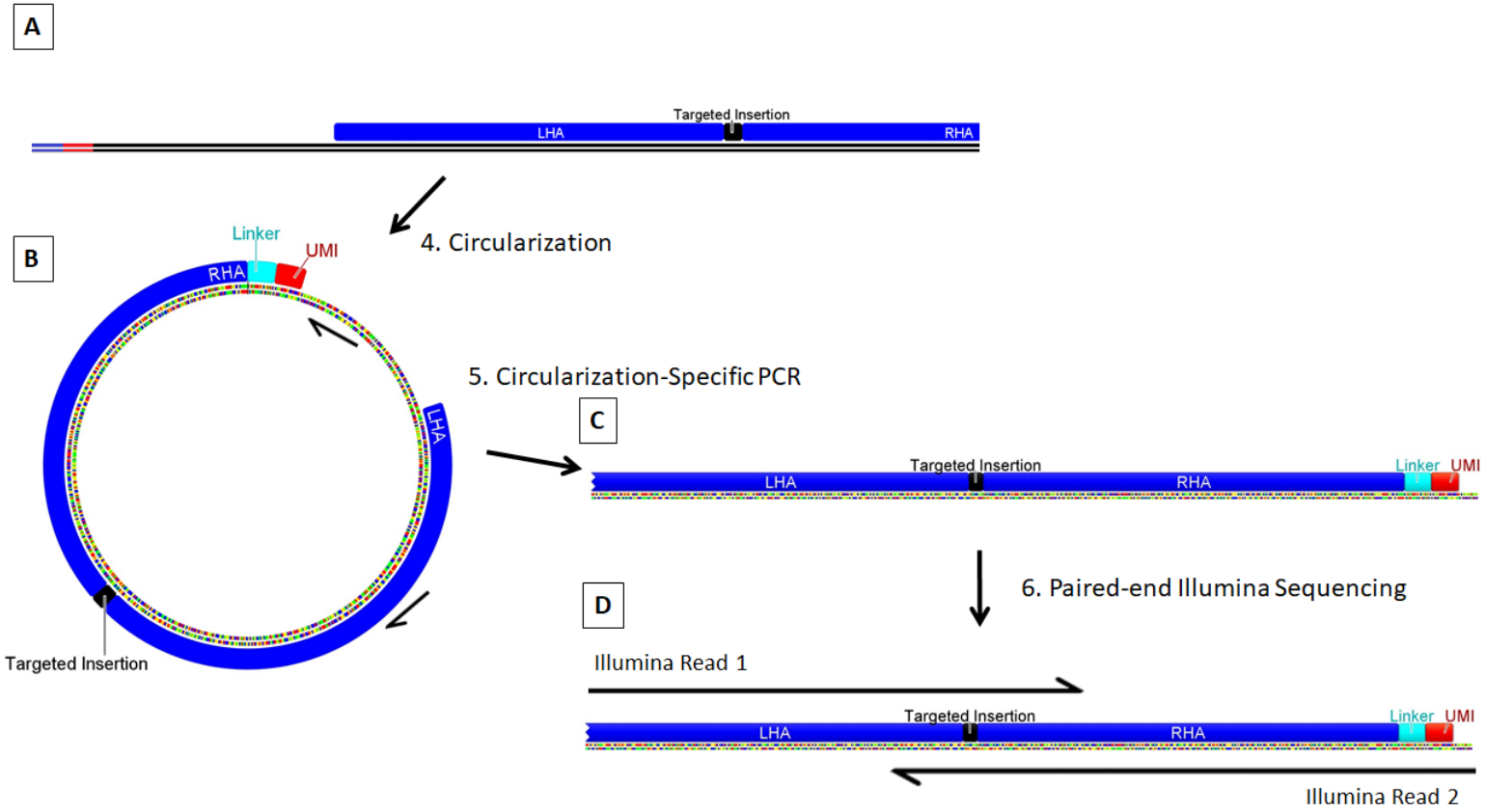

A
Targeted Insertion
LHA
RHA
4. Circularization
B
Linker
UMI
RHA
LHA
Targeted Insertion
5. Circularization-Specific PCR
C
Targeted Insertion
LHA
RHA
Linker
UMI
6. Paired-end Illumina Sequencing
D
Illumina Read 1
Targeted Insertion
LHA
RHA
Linker
UMI
Illumina Read 2

Figure 2-3 UMI circularization strategy enables paired reads to sequence across target site. (A) Amplicon prepared as in Figure 2b, but instead of the standard 30-40 amplification cycles are used to convert the single stranded extension product to a double-stranded product. Additionally, these oligos contain linker regions with unique restriction enzyme sites that will create compatible sticky ends when digested. (B) Double-stranded product is circularized by sequential digest and ligation. (C) Circularization-specific primers are used to selectively amplify circularized product. This step transposes the UMI from its initial position to the other end of the amplicon and allows for the second oligo to be placed anywhere within the donor arm. (C) Amplicon product is prepared for sequencing using a standard pipeline. (D) NGS reads spanning the amplicon.

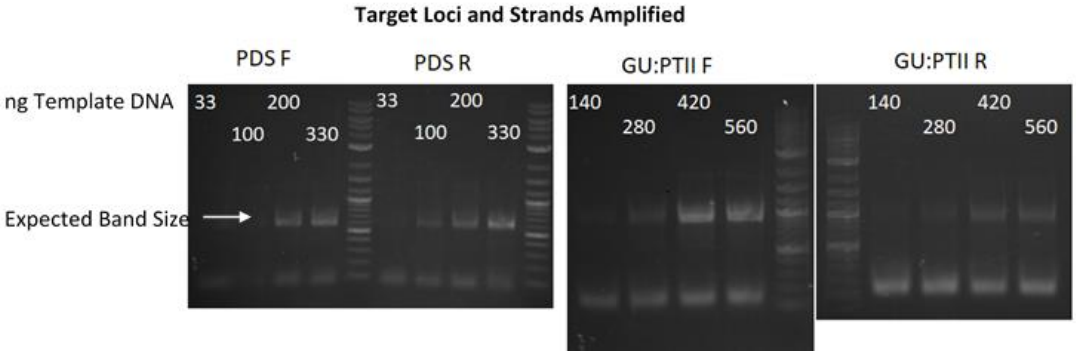


Figure 2-4 Optimization of linear extension conditions. Two targets (PDS in rice, GUP:TII transgene in *Nicotiana tabacum*) were amplified using the UMI sample preparation (Figure 2-4). F and R refer to forward and reverse strands (or Watson and Crick); each linear extension and subsequent analysis is strand-specific. The process was found to require significant quantities of genomic template DNA, often a limiting factor in protoplast experiments.

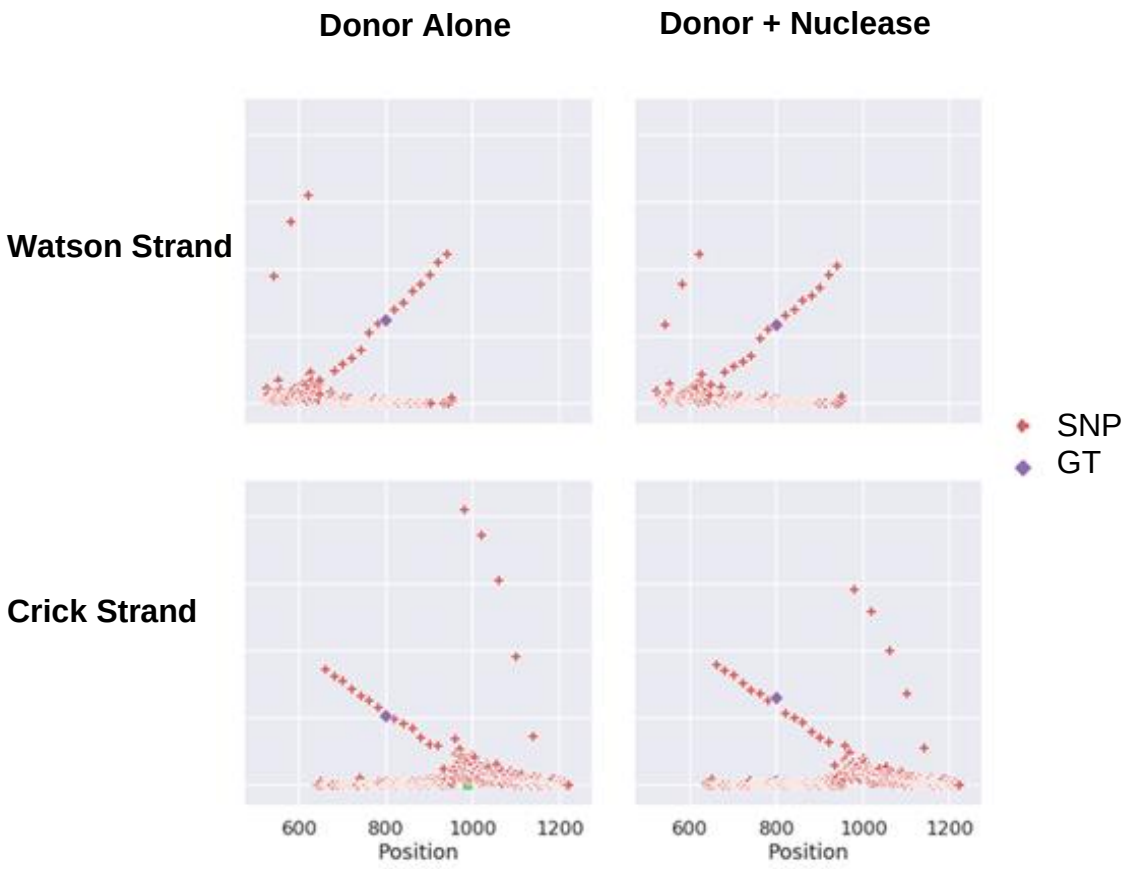


Figure 2-5 PANGEA analysis reveals rampant donor template switching in UMI circularization preparation. The Watson and Crick strands of the target region were separately analyzed in donor alone and donor plus nuclease (Cas9) treatments. Both were found to have nearly identical gene targeting frequencies and SNP deposition patterns. In this context, a nuclease treatment is expected to increase gene targeting by several orders of magnitude. Additionally, the reversal of the SNP deposition pattern upon analysis of the opposite strand heavily suggests template switching, facilitated by the binding of the donor region primer (Figure 2-2B) to donor molecules contaminating the genomic DNA sample and linear extension product even after purification. The directionality and decreasing frequency of SNPs is also consistent with this interpretation, as the template switching event is due to incomplete polymerase procession.

CHAPTER THREE

PANGEA, a Tool for Dissecting Genome Editing Outcomes with Nanopore Sequencing

Paul A.P. Atkins and Daniel F. Voytas.

Preface

Plant gene targeting outcomes are often to molecularly analyze due to residual editing reagents interfering with PCR. These limitations can be overcome by third generation sequencing, PacBio and Oxford Nanopore Sequencing (ONS), by sequencing large amplicons whose primers sit far outside the donor molecule. PacBio sequencing has been applied to analyze genome editing outcomes in several contexts, but it lacks ONS's ease to run highly multiplexed samples and low overhead. A minimal ONS experiment may be performed in under a day and for less than 250 USD, generating sufficient reads to quantify genome editing outcomes in dozens of samples. Despite this ease of access and speed, bioinformatic pipelines built to analyze ONS data from genome editing experiments do not exist. Here we describe a Phased Analysis of Gene Edited Amplicons, or PANGEA, a python-based bioinformatics analysis tool with a PYQT5-based GUI for quantifying genome editing outcomes and observing conversion tracts from an error-prone sequencing method.

Introduction

Oxford Nanopore Sequencing (ONS) is an emerging sequencing method with unique properties. Its low overhead, rapid turnaround time, and lack of read size restrictions make it a highly-flexible and accessible tool (Slatko, Gardner, and Ausubel 2018). This power brings with it a variable and context-dependent error rate that ranges from 8% to over 50% at some homopolymer sequences (Jain et al. 2018). A bioinformatic pipeline capable of accommodating this error rate would be highly valuable, allowing for rapid and cost-effective analysis of genome editing outcomes regardless of amplicon size. Here we present one such tool, PANGEA (Phased ANalysis of Gene Edited Amplicons), a Python-based program capable of quantifying genome editing outcomes as well as performing other analysis dependent upon long, phased reads produced by ONS (Figure 3-1A-B). Samples analyzed here were produced for previous work, and genomic DNA from these samples was generously provided by the Moriarity lab (Pomeroy et al. 2020).

Estimating targeted mutagenesis with PANGEA

PANGEA uses background subtraction to monitor and reduce the impact of sequencing error on estimates of targeted mutagenesis frequency from amplicons. This requires sequencing negative control amplicons (typically no nuclease) in parallel with the experimental sample (same library preparation and flow cell). PANGEA additionally requires the target reference sequence, nuclease target site, and donor molecule sequence. Experimental samples and a negative control sample are aligned to the reference using Minimap2, creating a SAM file whose CIGAR strings (Concise Idiosyncratic Gapped Alignment Report), a compact format encoding a read's sequence relative to a reference, are translated into a Pandas dataframe (H. Li 2018, 2). The dataframe consists of all mutations relative to the reference sequence (insertions, deletions, and single nucleotide polymorphisms (SNPs)) within a window of the target specified by the user. The mutation frequencies found within the same window in the control

sample are treated as the background error created during ONS and used in downstream processing.

The consistent nature of the error allows its subtraction from a control sample on a mutation-specific basis (Figure 3-2A-F). Mutation-specific subtraction allows for treatment-specific mutations to stand out more clearly (Figure 3-2D, 3-2F). One outcome of subtraction can be negative values for mutation frequency. This can be interpreted as both the error in the measurements and, when mutations at the target site in a nuclease-treated sample are particularly high, the loss of the WT sequence that was predisposed to a specific erroneous reading (homopolymers are ideal substrates for deletions created by microhomology-mediated end joining and also ONS error). For example, consider a homopolymer sequence at the target site containing an error in 50% of amplicons; if 25% of these homopolymers are disrupted by the nuclease treatment by a 1bp insertion, the amount of background will appear to have been reduced, but instead the prevalence of the error-prone sequence itself was reduced prior to sequencing, resulting in a negative mutation frequency for that particular background mutation. A more sophisticated background subtraction technique may be needed to precisely resolve this issue.

‘Fuzzy’ Gene Targeting Search

PANGEA’s primary purpose is the quantification and analysis of reads containing gene targeting events. A key aspect of PANGEA is its accommodation of sequencing error – how can a specific mutation at an exact position be detected if all reads can be expected to contain numerous sequencing errors? This makes searching for an exact sequence at an exact position ineffective and requires flexible search parameters. To this end, when searching for small gene targeting insertions, PANGEA accommodates sequencing errors that may alter the nucleotide sequence, size, and position of insertion by performing a ‘fuzzy’ search. This utilizes two ‘fuzz’ parameters – Levenshtein distance and position variance. In the context of nucleotides, Levenstein distance is the minimum number of alterations that can change one sequence into another via base

changes, insertions, or deletions (e.g. AATCG is 2 units diverged from ATCC, one deletion and one base change). Position variance is the variance in the alignment position of the gene targeting modification; errors at the junction of the gene targeting event and endogenous sequence may shift the alignment one or more bases.

The stochastic and context-dependent nature of sequencing error makes it difficult to predict error patterns *a priori*, and therefore the maximum amount of Levenshtein distance and position variance that may be accommodated without creating false-positives should be empirically tested. To facilitate this, PANGEA tests a matrix of fuzz parameters, repeatedly searching a nuclease only (non-donor) sample for gene targeting events with each combination (empirical_fuzz). The search ceases when gene targeting levels are found to have increased beyond an input threshold (default 0.01%). The output is a matrix of Levenshtein and sequence distances and their associated gene targeting frequency, allowing for the user to select the proper level of the fuzz parameters for analysis of their samples. The recommended fuzz levels for use on experimental samples are those with at most 0-2 gene targeting events or less than 0.01% gene targeting, keeping the Levenshtein distance below 50% of the insertions size (<9 if 18bp insertion) and the sequence distance to below 5 if examining targeted insertions. The empirical fuzz output will likely show values exceeding these without background, but these conservative recommendations are generous for all but the most error-prone targeted insertions (e.g. long homopolymer strings). This empirical fuzz test may be repeated on experimental samples containing donors, in this case a sharp increase in gene targeting events is expected as fuzz parameters initially increase (particularly the Levenshtein distance), which plateaus until they begin increasing again – the desirable settings are the smallest values at the first plateau. The second increase is expected to be the same fuzz parameters that result in detection of gene targeting events in the nuclease-only control. Settings should be such that minimal false positives will

occur in samples containing donors, establishing a sound baseline to which optimization experiments can be compared.

One alternative method is recommended for finding target insertions over ~50bp (tested on 1.4kb insertions) searches for mutations at or near the target site above an input size threshold (~80-90% of expected size). Insertions of that size are mapped to the donor insertion. Those that successfully align and pass a similarity threshold (percentage similarity or alignment score) count as gene targeting events. Targeted deletions are detected in a similar manner, although this depends upon the user to deliver a donor with insertions larger or somehow distinct from those created by NHEJ/MMEJ deletions. Deletions above an input size threshold and within an input range of the target site are counted as gene targeting events.

Mechanistic Analysis

Phased analysis is made possible by SAM file processing; mutation information is stored on a per-read basis. This allows reads containing gene targeting events to be searched for associated mutations and for those gene targeting-associated mutations can be compared to unmodified reads. Three experimental contexts for which this is relevant will be addressed here – mechanistic analysis using SNP-laden donor arms, gene conversion experiments, and detecting PCR template switching.

Historically, mechanisms of non-meiotic gene targeting and homologous recombination have been addressed by tracking the deposition of SNPs copied from a donor molecule into the genome (Szostak et al. 1983; Heyer, Ehmsen, and Liu 2010; Symington and Gautier 2011; Kan et al. 2014). This has been performed in yeast and mammalian cells by delivering a SNP-laden donor which, upon precise insertion into the genome, results in antibiotic resistance, enabling clonal propagation and analysis of edited cells. In these experiments, distinct SNP patterns are attributed to distinct mechanisms of recombination, with bidirectional incorporation of SNPs indicating a traditional double Holliday

junction intermediate pathway and unidirectional incorporation indicating synthesis-dependent strand annealing (SDSA). More recent studies have implicated other pathways (primarily single-strand assimilation) and suggested that an SDSA mechanism may still incorporate SNPs bidirectionally due to limited 3' resection (Kan and Hendrickson 2019). In plants, mechanistic analysis has been performed over nearly three decades, but with few conclusive results (H. Puchta, Dujon, and Hohn 1996; Holger Puchta 1998; Lieberman-Lazarovich et al. 2013; Schmidt, Pacher, and Puchta 2019a). Careful observation of SNP deposition patterns from many gene targeting events can determine the underlying mechanisms, but with the caveat that those SNP-containing events will be at a much lower frequency than perfect homology would otherwise produce (Opperman, Emmanuel, and Levy 2004; Emmanuel, Yehuda, Melamed-Bessudo, et al. 2006; L. Li, Jean, and Belzile 2006; Gonzalez and Spampinato 2020).

PANGEA can be used to analyze SNP deposition patterns; a user may input the donor molecule sequence that encodes SNPs of interest. These SNP positions are tracked in both unmodified and gene targeting-positive reads and extracted for further analysis. When investigating mechanism, the outermost SNPs (the furthest extent of copying) are most relevant, but the interior SNPs can be used to infer if the outermost SNP is a sequencing error. This is due to proposed copying mechanisms being continuous, making single gaps in a string of SNPs probable sequencing error and, conversely, any lone SNPs probable sequencing error as well. These assumptions allow for a simple, conservative approach to removing spurious outmost SNP that may confound results – any outmost SNP with four consecutive WT inner SNPs are ignored (see Chapter 4). This same SNP analysis can be used to track both gene conversion and PCR template switching (see Chapter 2). In both cases the noise reduction serves to draw attention to reads containing a substantial range of copied information.

Examining the Origin of Nuclease-induced Insertions

Another question surrounding genome editing is the frequency and nature of insertions at a nuclease target site other than the desired outcomes. Long-read sequencing gives a much greater opportunity to examine these events due to larger amplicons sizes accommodating a larger size range of insertions without them being lost during PCR cycling. To address this using PANGEA, a user may input a range surrounding the target site and a minimum size to examine insertions that meet those criteria. An additional input for this analysis is an ordered list of sequences that are potential origins for the target site, beginning with the delivered genome editing reagents (donor molecule (if applicable), reagents maps) and other relevant sequences (target site, broad target region, etc.). Insertions matching the specified criteria will be aligned to the first sequence in the ordered list. Any mapped sequences will be subtracted from the population of insertions and designated as originating from that sequence. The remaining insertions will then be mapped to the second sequence in the ordered list, with those mapping again being removed from the population of insertions and designated as originating from the second sequence. This will continue until the ordered list has been exhausted and will result in FASTA files corresponding to each origin containing the relevant reads and one additional FASTA file containing the remaining unmapped sequences that may be further analyzed. Relative and absolute prevalence of the insertions may be visualized in multiple way and gives a broad view of the nature of insertions found at the target site (Figure 3-3A-C).

Conclusions

PANGEA enables the analysis of genome editing experiments using ONS. For many labs, this would enable analysis of large experiments in 1-2 days with only a few thousand dollars of capital investment, a fraction of the cost of minimal Illumina or PacBio setups. ONS long reads allows for additional analyses requiring long, phased reads, particularly the ability to find large gene targeting insertions, particularly the discovery and classification of unexpected insertions at

the target site and tracking SNP deposition patterns for mechanistic analysis. The error-prone nature of ONS was found to be minimally impactful – conservative filters and background subtraction decreased the impact of sequencing error on frequencies of targeted mutagenesis and gene targeting. PANGEA makes it possible to perform and deeply analyze genome editing experiments with extreme speed and low cost.

Acknowledgements
Special thanks to the Moriarity lab for providing the samples for early testing of the PANGEA pipeline.

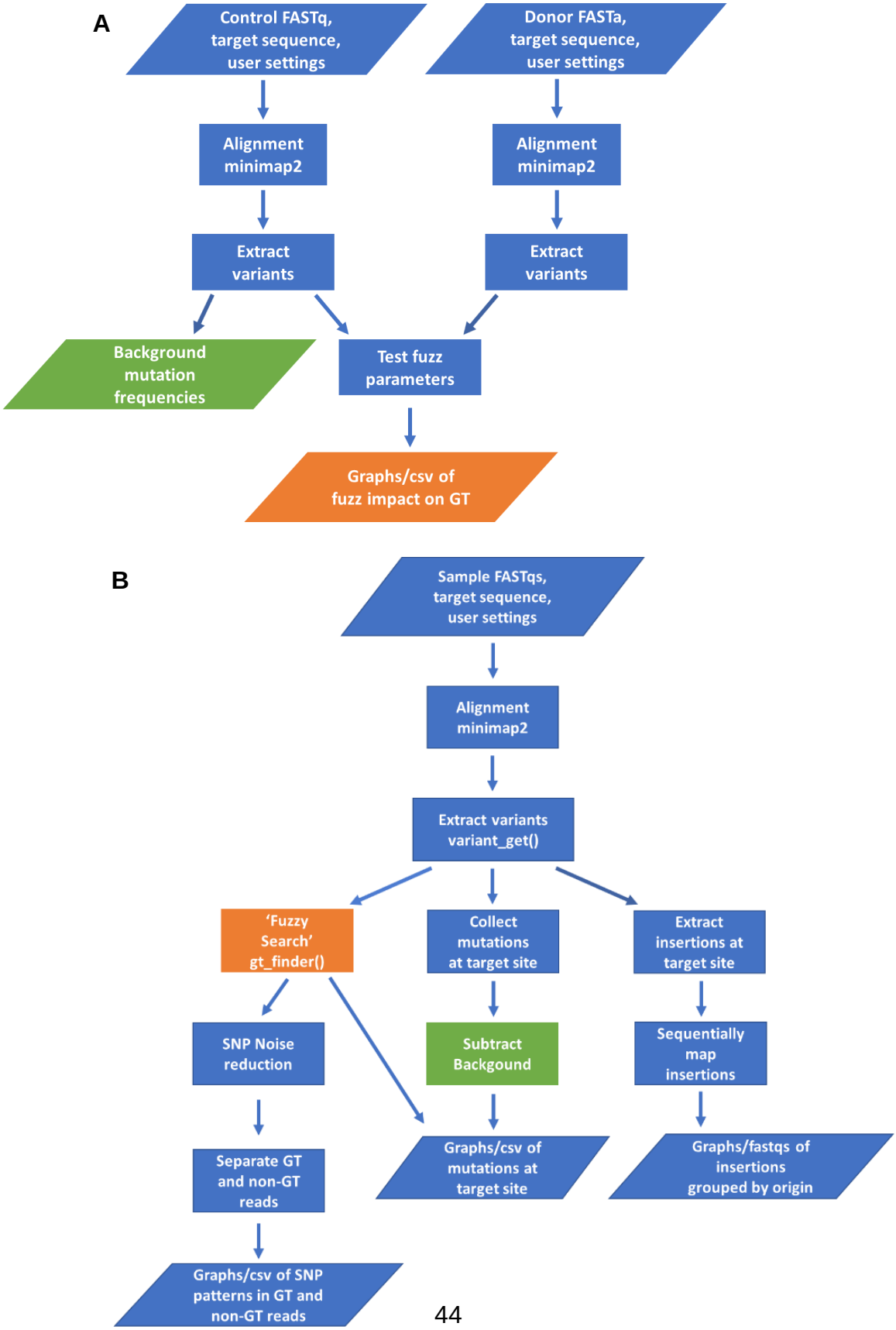
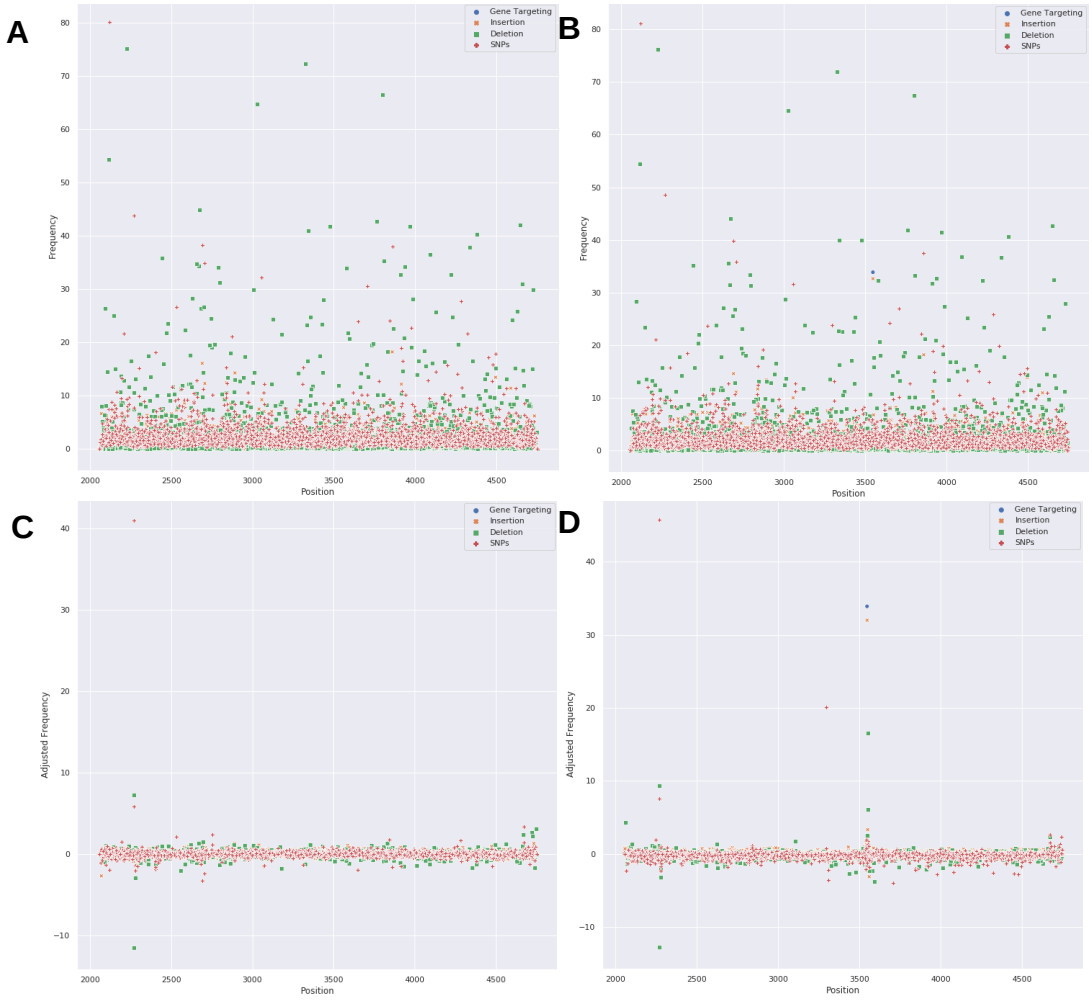

A
Control FASTq, target sequence, user settings
Donor FASTa, target sequence, user settings
Alignment minimap2
Alignment minimap2
Extract variants
Extract variants
Background mutation frequencies
Test fuzz parameters
Graphs/csv of fuzz impact on GT
B
Sample FASTqs, target sequence, user settings
Alignment minimap2
Extract variants variant_get()
'Fuzzy Search' gt_finder()
Collect mutations at target site
Extract insertions at target site
SNP Noise reduction
Subtract Backgound
Sequentially map insertions
Separate GT and non-GT reads
Graphs/csv of mutations at target site
Graphs/fastqs of insertions grouped by origin
Graphs/csv of SNP patterns in GT and non-GT reads

Figure 3-1 Flowcharts depicting PANGEA data processing. (A) **Left** – Initial analysis of control sample and inputs necessary for background subtraction and fuzz parameter determination. Fastqs from a control sample are aligned to the target using minimap2. All variants in the resulting SAM file are extracted and the frequency of all individual mutations at all positions are recorded. This serves as the background frequencies for subtraction. **Right** – Donor sequences are aligned to the target and all mutations are recorded. The largest insertion is assumed to be the desired gene targeting modification and the target site is assumed to be its point of insertion unless otherwise specified in the user settings. This gene targeting event is then searched for in all the control variants using a variety of fuzz parameters. The resulting gene targeting frequencies are output as a csv and may be readily visualized as needed. (B) Analysis of experimental samples. Each sample is aligned to the target using minimap2 and variants are extracted from the SAM file into a readily iterable format. **Left** – reads are examined for presence of gene targeting events using the fuzz parameters determined previously (orange boxes). SNPs matching those found in the donor are extracted from all reads. Reads are separated into GT and non-GT reads. SNP patterns may be clarified using the approach discussed in the text. SNP patterns for GT and non-GT reads are graphed for all samples. **Middle** – All mutations at the target site (default a 7bp region with the predicted cut base at the center) are recorded and the background frequency of each mutation is subtracted by its frequency in a control sample (green) to determine the frequency of targeted mutagenesis. **Right** – All insertions at the target site above an input threshold size (default 10bp) are extracted and sequentially mapped to potential sources, as described in the text. Outcomes from the analysis are then graphed.


A
B
C
D
Gene Targeting
Insertion
Deletion
SNPs
Frequency
Adjusted Frequency
Position

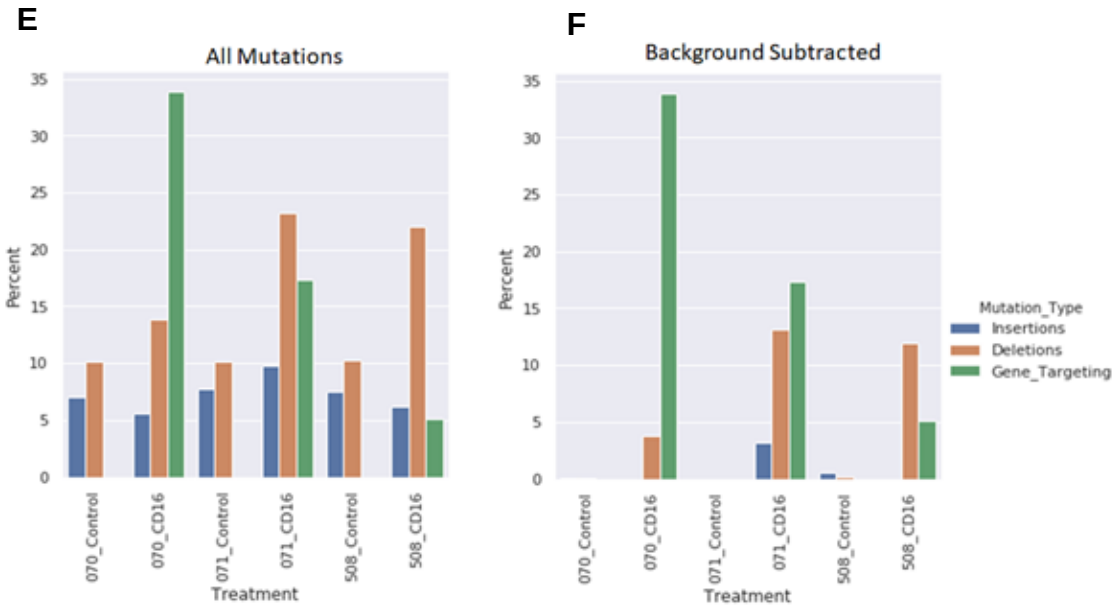

E
All Mutations
F
Background Subtracted
Percent
Treatment
070_Control
070_CD16
071_Control
071_CD16
508_Control
508_CD16
Mutation_Type
Insertions
Deletions
Gene_Targeting

Figure 3-2 Background subtraction approach highlights nuclease treatment-specific modification. Frequency and distribution of insertions, deletions, SNPs, and gene targeting events in PANGEA-analyzed amplicons from cells delivered either a non-nuclease control or GT reagents inserting a CD16 sequence at the target site. CD16-labeled samples include a nuclease and donor molecule. (A) Non-nuclease treatment, showing the variability of nanopore sequencing error. (B) Cells treated with gene targeting reagents (C) Frequencies in A after subtraction from another negative control sample prepared in parallel. (D) Frequencies in B after subtraction from a negative control revealing prominent reagent-specific mutations at the target site. (E) Percentage of insertions, deletions and gene targeting events at the expected target site for 3 GT-reagent treated and 3 control samples. (F) As in E but post background subtracting using 070_Control for all samples.

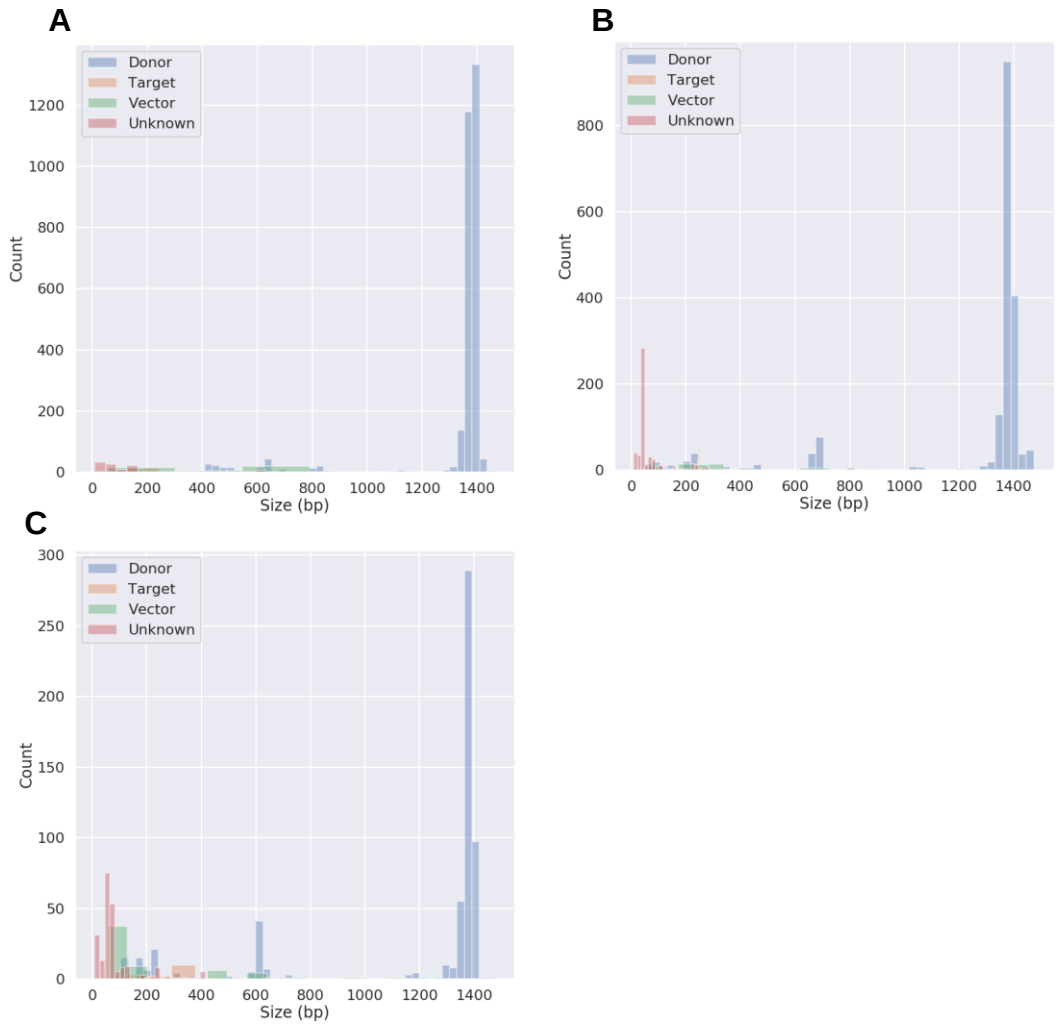


Figure 3-3 Vast majority of targeted insertions are of the desired donor sequence. Histograms of sizes of targeted insertions found within three samples delivered Cas9 and donor molecules. Most targeted insertions above 10bp were found to be of the expected donor molecule at approximately the expected size (1400bp). In all samples, most small insertions were not found to originate from any vector components or sequences nearby the target. (A) Sample 070. (B) Sample 071 (C) Sample 508.

CHAPTER FOUR

Analysis of Somatic Plant Gene Targeting Events Using Nanopore Sequencing

Paul A.P. Atkins, Maria Elena Gamo, and Daniel F. Voytas.

Preface

Efficient gene targeting tools have been sought by the plant community for nearly 30 years. Despite this, protocols that would allow a researcher to perform gene targeting experiments at a non-selectable/screenable target of interest are virtually nonexistent (Huang and Puchta 2019). This is largely the result of the low throughput of plant gene targeting experiments and the historic reliance on gene targeting reporters at biased targets. Here we address these issues by performing a target agnostic, molecular analysis of gene targeting outcomes in somatic *Nicotiana benthamiana* using Oxford Nanopore amplicon sequencing. Our results reveal the low frequency of somatic plant gene targeting events using 500bp homology arms, the extreme sensitivity of gene targeting to sequence variation in the donor arms, and the small size and patterns of conversion tracks copied during somatic homologous recombination. Further, we have established a readily scalable framework for optimization and study of gene targeting that may be applied to any organism.

Introduction

The ability to precisely modify any sequence within a plant genome would massively accelerate basic and applied plant research. Gene Targeting (GT) is one prominent method to this end, wherein a genomic sequence is repaired by homologous recombination using an engineered donor molecule as a template (Thomas and Capecchi 1987). The efficiency of gene targeting has been found to be extremely variable between organisms and to be uniformly enhanced by a targeted double-strand break (DSB) at the modification site (Hohn and Puchta 1999). Analysis of plant gene targeting outcomes have been largely restricted to phenotypic reporters that utilize a screen or selection to identify positive events (Holger Puchta and Fauser 2015). This has prevented the use of gene targeting tools for loci of general research interest, restricting it to a handful of endogenous targets or entirely synthetic integrated reporters.

Genome editing outcomes are difficult to measure in the presence of large donor molecules, which is the standard in plants. These donors interfere with high-throughput molecular analysis methods, particularly Illumina sequencing and qPCR-based approaches, due to the inability to place primers within the large donor arms. Third generation sequencing techniques allow researchers to sequence long amplicons, bypassing size restrictions (Hendel et al. 2014; Canaj et al. 2019). Here we utilize Oxford Nanopore Sequencing (ONS) to assess genome editing outcomes in plants and develop tools for overcoming its primary limitation – sequencing error.

We found that the error rate of Nanopore sequencing can be largely mitigated when estimating frequencies of targeted mutagenesis and finding GT-positive amplicons. As expected, we found that GT efficiencies were low using standard Cas9 reagents and 500bp homology arms and were enhanced when the donor molecule and gRNA cassette were replicated by geminiviral replicons (GVRs), a ssDNA virus-based bioreactor repurposed for genome editing (Baltes et al. 2014).

Analysis of conversion tracts is facilitated by a donor with perfect homology to only one homologue and imperfect homology to the other. This allowed us to monitor gene targeting outcomes at both a perfect and imperfect target for each treatment (deliver donor molecule with perfect homology to either PDS3.1 or PDS3.2 and sequence both PDS3.1 and PDS3.2). We found conversion tract patterns consistent with both synthesis-dependent strand annealing (SDSA) and double Holliday junction (dHJ) mechanisms and that repair using a donor with divergent sequence is severely reduced, as previously reported (Opperman, Emmanuel, and Levy 2004; Emmanuel, Yehuda, Melamed-Bessudo, et al. 2006, 2; Gonzalez and Spampinato 2020). Additionally, we found that most gene targeting events captured here utilized short conversion tracts, which may have significant implications for future gene targeting optimization.

Results

Measuring genome editing outcomes with Nanopore Sequencing requires accommodation of its high error. To determine the feasibility of this, T-DNAs encoding genome editing reagents were delivered via leaf infiltration to *Nicotiana benthamiana* encoding a 35S-driven Cas9 transgene. These reagents consisted of a single gRNA that targets two genomic PDS homologues (hereafter referred to as PDS3.1 and PDS3.2), and one of two possible donor molecules. Each donor has 500bp homology arms designed to insert an 18bp sequence precisely at the gRNA cut site, each possessing perfect homology to either PDS3.1 or PDS3.2 (6% sequence divergence, Figure 4-1A). Additionally, a subset of these T-DNAs encoded a GVR. 5 days after delivery, DNA was extracted and allele-specific PCRs were performed on all samples. The resulting 1650bp and 1594bp amplicons (PDS3.1 and PDS3.2, respectively) were sequenced by ONS and demultiplexed using minibar (Krehenwinkel et al. 2019). The resulting fastq files were aligned to their respective templates using minimap2 and assessed for insertions, deletions, and SNPs at all positions of the amplicon (Supplemental Figure 4-1A-D) (H. Li 2018). The sequencing error rate at the target site (defined as 7bp window surrounding the predicted nuclease cut site) ranged between 25

and 36% (Supplemental Figure 4-2). We observed that, except at the target site, nuclease and non-nuclease treatments possessed a consistent error pattern (Supplemental Figure 4-3A-B). Based on this, a subtraction strategy was used to estimate frequencies of targeted mutagenesis; mutations found within a non-nuclease control sample were subtracted from treated samples on a mutation-specific basis. This facilitated the identification and quantification of nuclease-specific mutations (Supplemental Figures 4-1E-H and 4-3C-F). This approach may result in inaccurate estimates of targeted mutagenesis if nanopore sequencing error and genome editing share common events, but these estimates may be supplemented by other approaches such as ICE or TIDE for validation at new targets.

Nanopore sequencing's high error obscures gene targeting modifications, making searches for the exact, desired sequence modification unreliable. To address this, we incorporated two 'fuzz' parameters into the search for gene targeting events to accommodate error: Levenshtein distance and sequence distance. Levenshtein distance accommodates base changes from sequencing error while sequence distance accommodates changes in alignment position. High 'fuzz' parameters gave rise to virtually no false positives and values were chosen that appeared to maximize the number of gene targeting events while maintaining zero false positives in a nuclease only control (Supplemental Figures 4-4 and 4-5). All further analyses utilize error-corrected frequencies of targeted mutagenesis and gene targeting events found with the fuzz parameters described in the supplemental material.

Genome editing outcomes were compared at PDS3.1 and PDS3.2 after delivery of gRNA and donors with 500bp homology arm via Agroinfiltration to *Nicotiana benthamiana* leaves constitutively expressing Cas9. *PDS3.2* showed higher efficiencies of targeted mutagenesis compared to PDS3.1 and efficiencies at both loci were increased by the GVR (Figure 1B). Gene targeting frequencies were not found to be significantly different at the two loci, although sampling was

limited (Figure 4-1C). Donors with perfect homology arms containing no mutations were significantly more efficient than donors with imperfect homology and these differences appear to be exacerbated by the GVR (Figure 4-1D) whereas differences in targeted mutagenesis were unaffected (Supplemental Figure 4-6).

Next we examined the impact of a set of donor variants (inverted donor and single homology-arm donors) on gene targeting efficiency at PDS3.1 (Figure 4-1E). Single-arm donors may inform the mechanism by allowing for strand invasion from only one direction; a strong bias in repair frequency or differences in conversion tracts between single-arm donors would indicate these preferences. An inverted donor was delivered to determine if ssDNA produced during GVR rolling circle replication (RCR) of GVRs was contributing to its high gene targeting frequency. Inverting the donor switches the strand most prevalent during RCR, and if a ssDNA template is the preferred GT template, conversion tract patterns will invert when the donor is inverted (see below). Single-arm donors were found to be extremely inefficient (Figure 4-1E). The inverted donor was found to have no impact on the frequency of gene targeting when delivered with both standard and GVR T-DNAs, suggesting a dsDNA and not a ssDNA substrate is preferred as a GT template (Figure 4-1E).

Donor-specific SNPs found in gene targeting events copied from imperfect donors were tracked at both targets to determine the conversion tracts and mechanisms used during gene targeting. To minimize error due to random mutations introduced by sequencing, only SNPs with less than three consecutive internal non-SNPs were counted (Supplemental Figure 4-7). When these criteria are applied to non-GT reads, we see a significant reduction in the background resulting from ONS error at all but the most internal SNPs (Supplemental Figure 4-8). We found that most GT events in all treatments had short (under 100bp) conversion tracts (Figure 4-2A-F). GVR treatments tended to have longer conversion tracts and more pronounced directional biases, incorporating more

distal SNPs from the right homology arm. The inverted donor maintains this trend, suggesting its effect on conversion tracts is due to the pleiotropic effects of the GVR replicase rather than uneven production of the Watson and Crick strands during rolling circle replication.

In order to draw mechanistic conclusions, SNPs patterns were classified as being bidirectional (incorporated SNPs from both sides of target), unidirectional (SNPs from one side of target), or having no SNPs detected (Figure 4-3). Interestingly, both bidirectional and unidirectional SNP patterns were found in all treatments but with high variation within and between treatments. This was also performed on non-GT reads in order to observe the impact of ONS error on conversion tract patterns. Non-GT reads were found to have extremely rare conversion tracts consistent with nanopore error and the noise reduction methods taken (Supplemental Figure 4-7). The extremely low level of bidirectional conversion tracts found due to their low frequency limits our interpretation of trends.

Discussion

Here we assessed analyzed genome editing outcomes in *Nicotiana benthamiana* leaf tissue with ONS. We found that the high error rate of ONS could be accommodated when determining frequencies of both targeted mutagenesis and gene targeting. The former was possible because sequencing error was largely consistent between amplicons sequenced on the same flow cell. This led us to subtract the errors found in the negative control from the treatments to reveal treatment-specific mutations. At some targets, error subtraction may be confounded when a consistent indel sequencing error (as is expected at homopolymers) is shared by a common targeted mutagenesis outcome, or alternatively when the common outcome creates a sequence prone to error. Overlapping errors between genome editing outcomes and ONS make this strategy not ideal for precise comparisons of targeted mutagenesis outcomes. Instead, it serves to estimate frequencies and may be improved by a

deeper understanding of target-specific error patterns and supplemented by other estimation methods (e.g. ICE or TIDE)

Unlike outcomes of targeted mutagenesis, gene targeting events are readily distinguishable from sequencing error with minimal accommodation. For the small (18bp) insertions assessed here, we found searching for sequences within 6 Levenshtein distance and within 8bp of the expected target base resulted in no false positives in nuclease only treatments.

Gene targeting frequencies were found to be extremely low when reagents were delivered on a standard T-DNA and drastically enhanced by the GVR at both PDS homologues. This trend did not hold for non-standard (500bp homology arm) donors: both the removal of one donor arm and the presence of mismatches resulted in low frequencies of gene targeting that were unchanged by the addition of the GVR. This lack of improvement may be due to these low frequency events being outside the linear dynamic range of PCR or that GVR-treated cells are particularly sensitive to donor mismatches. In either case, these low efficiencies cannot be attributed to variations in delivery or reagent efficiency as the frequencies of targeted mutagenesis were consistent within GVR and non-GVR treatment groups. These results highlight the technical challenges in molecular characterization of somatic plant gene targeting events.

Conversion tract analysis revealed a combination of unidirectional and bidirectional repair events in all gene targeting treatments, suggesting repair by both SDSA and dHJ pathways. The high variance in these patterns may be due to their low frequency and the subsequent poor sampling of the events, but the variety of outcomes and the pooling of 3 independent amplifications for each sample suggest the results may be biologically relevant.

These data demonstrate a high-throughput nanopore-based pipeline for assessing genome editing efficiencies not restricted by reagent design that may be applied to virtually any target. We found plant gene targeting frequencies to

be extremely low and greatly increased by GVRs, and that GT occurred using a combination of dHJ and SDSA pathways. This work serves as a blueprint for future high-throughput gene targeting experiments in any organism using ONS.

Materials and Methods

Vector Construction: T-DNA plasmids were constructed using a previously published Golden Gate system for plant genome engineering. Intermediate vectors and amplicons used to create the final T-DNA vectors are listed in Supplemental Tables 1 and 2. PDS gRNA golden gate construct described previously (Maher et al. 2020). PCR templates encoding *PDS3.1* and *PDS3.1* donor molecules were synthesized by Twist Bioscience.

Plant Material: *Nicotiana benthamiana* encoding a transgene expressing Cas9 under the 35S promoter has been previously described. These plants were grown at 24C and 60% humidity, 16h/8h day/night cycle in a Conviron growth chamber. Plants were selected for infiltration after 4-5 weeks of growth.

Leaf Infiltration and DNA Isolation: GV3101 *Agrobacterium* were transformed via freeze thaw method and delivered to true leaves via leaf infiltration as previously described.(Baltes et al. 2014) DNA was extracted from infiltrated plants 5 days post infiltration using a modified, plate-based CTAB method.(Cody, Graham, and Birchler 2017)

PCR Amplification of Genome Editing Outcomes: Primers and barcodes used for PCR amplification of *PDS3.1* and *PDS3.2* are described in Table 3. Amplifications were performed using PrimeStar GXL Polymerase using the manufacturer's guidelines. 3 25ul PCRs were performed for each sample at both *PDS3.1* and *PDS3.2*. Approximately 2ul of each sample was examined on an agarose gel to verify successful amplification. The 3 PCR replicates were then pooled and purified using a single Qiagen PCR Purification Kit. DNA concentrations of the purified amplicons were determined using Nanodrop and equimolarly pooled.

Nanopore Library Preparation and Sequencing: Sample amplicon pool underwent library preparation using a SQK-LSK109 Oxford Nanopore Ligation Sequencing Kit, which was performed using the manufacturers specifications using the Short Fragment Buffer (SFB) protocol variant. The completed library was sequenced using a R9.4.1 MinION flow cell on a MinION device according to the manufacturer's specifications.

Basecalling, Demultiplexing, and additional bioinformatics: Fast5 files were converted to nucleotide base calls using Oxford Nanopore's Guppy Basecalling Software, version 3.3.3+fa743a6 using an RTX2080 TI and Ubuntu 18.04. The resulting fastq files were demultiplexed using Minibar.(Krehenwinkel et al. 2019) Additional bioinformatics were performed using Python 3.7 and graphed using Seaborn (https://seaborn.pydata.org/) and pandas (https://pandas.pydata.org/). Code used for analysis of genome editing outcomes in nanopore reads can be found at https://github.com/atkin265/PANGEA.

Author Contributions

PAPA and DFV conceived and planned the research. PAPA wrote the text, developed the bioinformatic tools, carried out the analysis, and performed experiments. MEG generated vectors and performed leaf infiltration experiments.

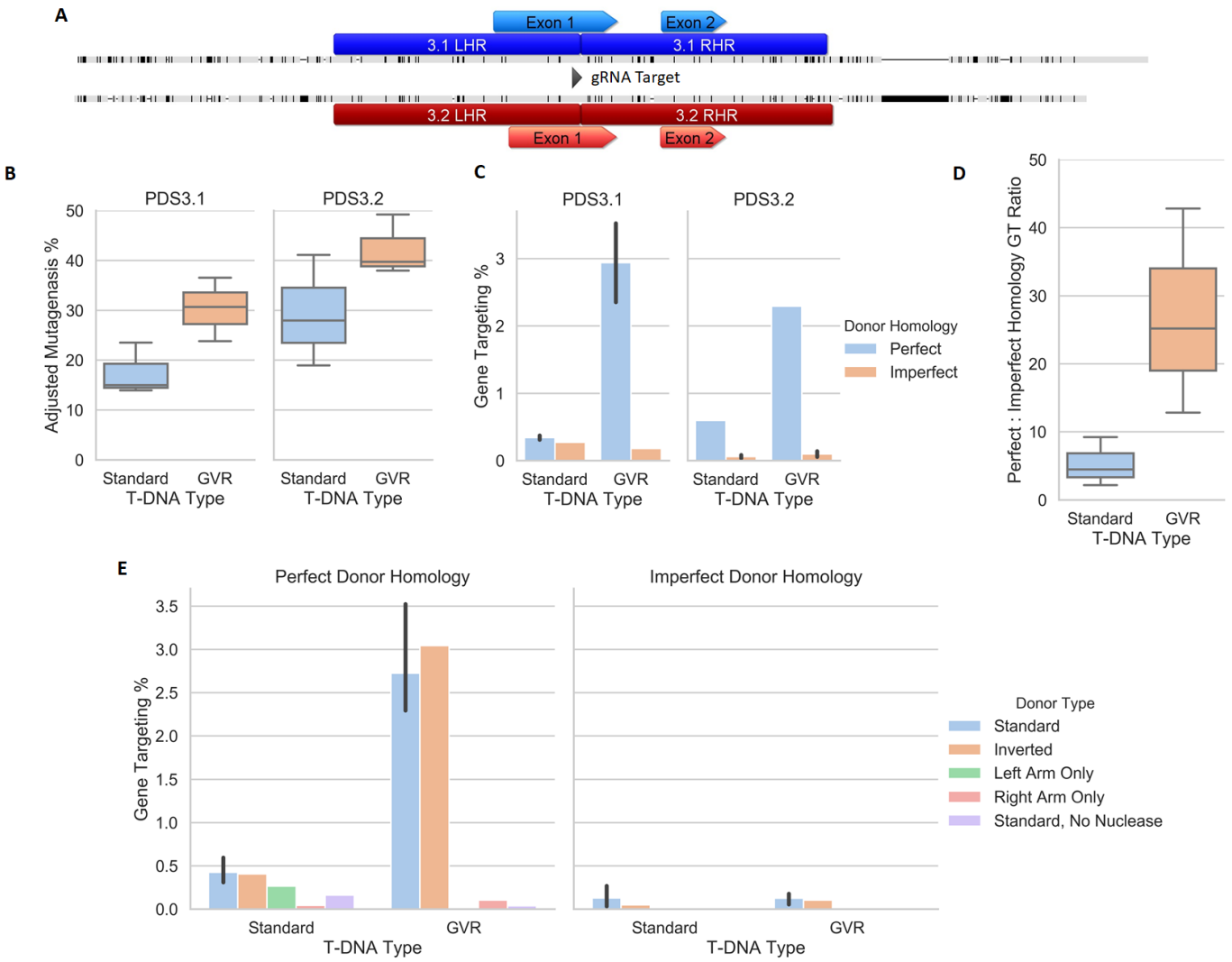

A
Exon 1
Exon 2
3.1 LHR
3.1 RHR
gRNA Target
3.2 LHR
3.2 RHR
Exon 1
Exon 2
B
PDS3.1
PDS3.2
Adjusted Mutagenesis %
Standard
GVR
T-DNA Type
C
PDS3.1
PDS3.2
Gene Targeting %
Donor Homology
Perfect
Imperfect
D
Perfect : Imperfect Homology GT Ratio
E
Perfect Donor Homology
Imperfect Donor Homology
Gene Targeting %
Donor Type
Standard
Inverted
Left Arm Only
Right Arm Only
Standard, No Nuclease

Figure 4-1 Schematic and outcomes assessed via ONS for genome editing experiments at *Nicotiana Benthamiana PDS3.1* and *PDS3.2*. (A) Schematic of target sites and donors. A gRNA targeting two PDS homologues in Benthi is delivered with a single donor molecule encoding perfect homology to either *PDS3.1* or *PDS3.2*. 5 days post-delivery, genomic DNA is isolated and homologue-specific PCRs are performed to analyze editing outcomes at both loci. (B) Adjusted targeted mutagenesis frequency at two *PDS* homologues. The GVR impacted mutagenesis frequencies at both homologues, with higher overall mutagenesis was observed at *PDS3.2*. (C) Gene targeting frequency at two *PDS* homologues is enhanced by the GVR, but only when using a perfectly homologous donor. Perfect donor homology is repair in treatments when the matching donor is delivered (*PDS3.1* GT repair with *PDS3.1* donor) and imperfect donor homology is repair with the homologue donor (*PDS3.1* GT repair with *PDS3.2* donor, and vice versa). Frequencies are low for all but the standard donor delivered with the GVR. Imperfect donor GT frequencies were not enhanced by the presence of the GVR. (D) Ratio of GT events between perfect and imperfect donors within individual treatments. Perfect and imperfect donors as described in 1C. (E) Left Panel: GT frequency using donor variants with perfect homology (*PDS3.1* and *PD3.2* data combined). Right Panel: As in left panel, but instead using imperfect homology donors.

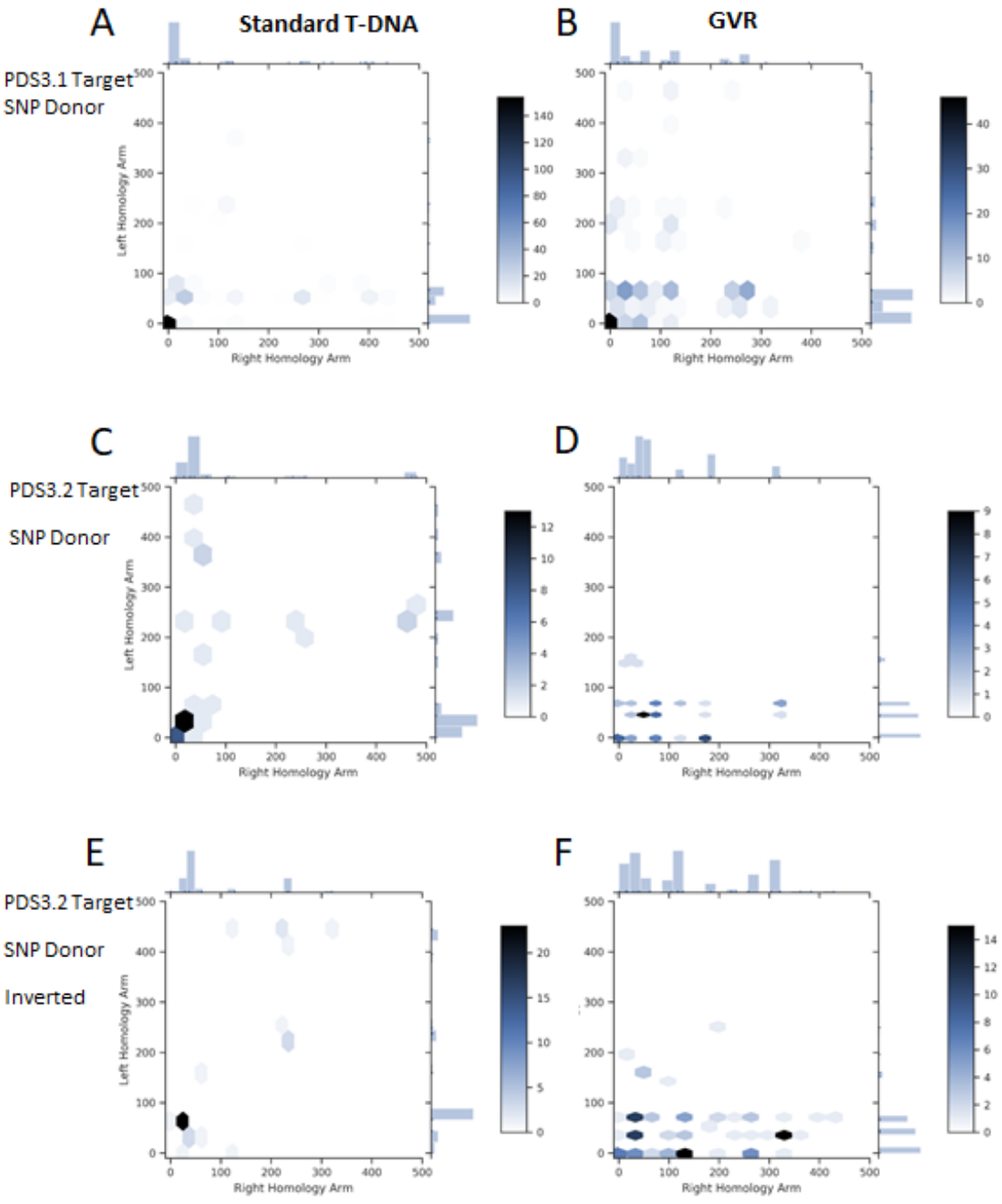

A
Standard T-DNA
B
GVR
PDS3.1 Target
SNP Donor
C
PDS3.2 Target
SNP Donor
D
E
PDS3.2 Target
SNP Donor
Inverted
F
Left Homology Arm
Right Homology Arm

Figure 4-2 Conversion tract patterns extracted from GT reads after noise reduction measures. (A) Conversion tract patterns at *PDS3.1* from sample treated with gRNA and *PDS3.2* donor taken from GT-positive reads. 265 events (0.27% GT). (B) As in A, but donor and gRNA were delivered using a GVR. 185 events (0.18% GT). (C) As in A, but with the target and donor reversed. 92 events (0.08% GT). (D) As in C, but reagents delivered using a GVR. 52 events (0.05%). (E) As in C, but the donor is in the opposite orientation within the vector (identical donor sequence). 40 events (0.05%). (F) As in E, but the donor and gRNA are delivered using a GVR. 114 events (0.10% GT).

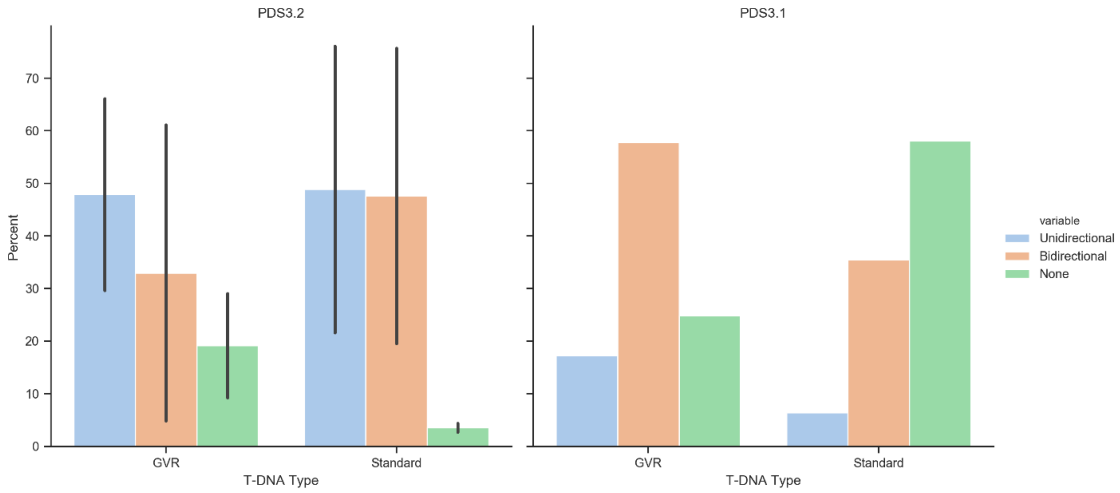
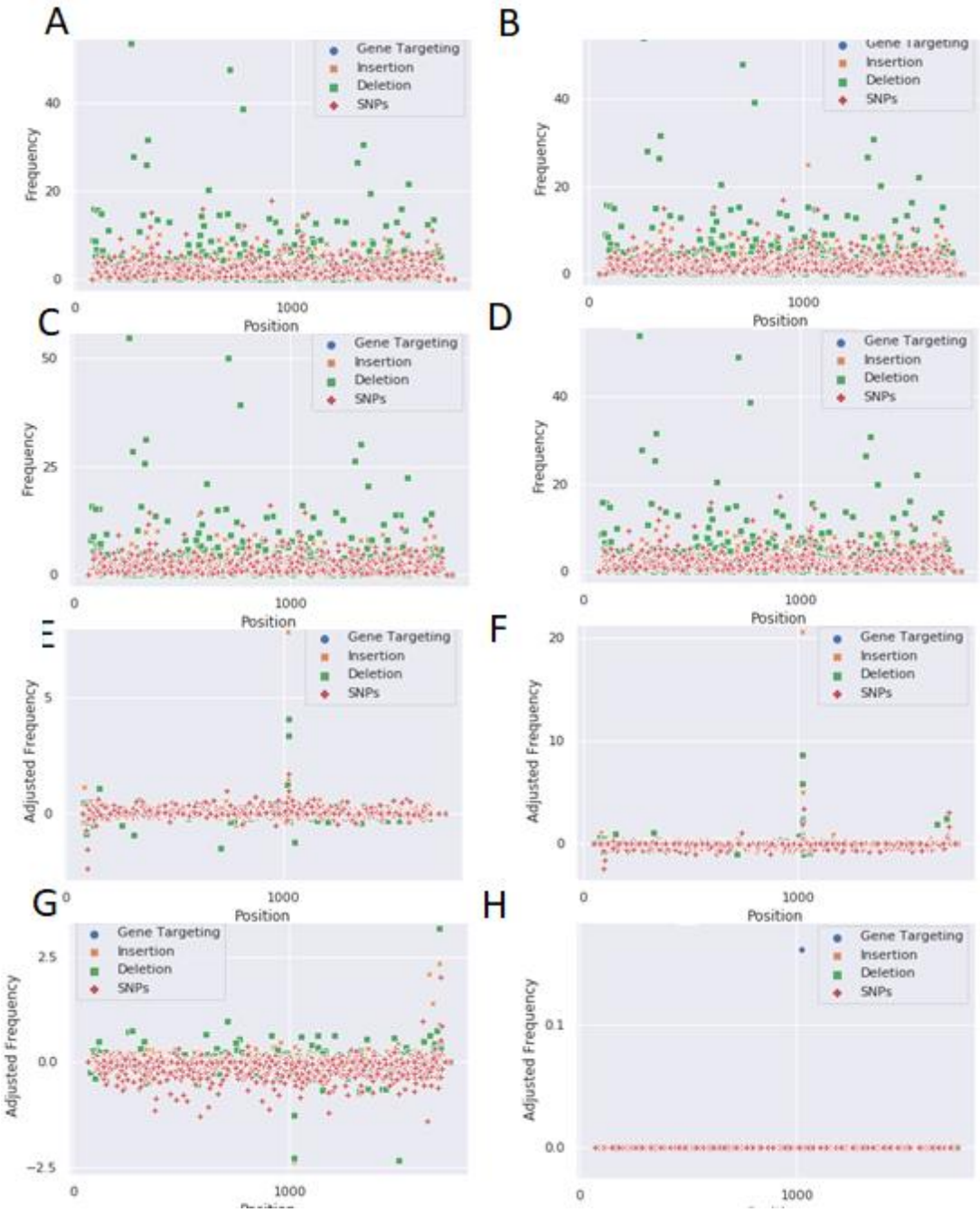
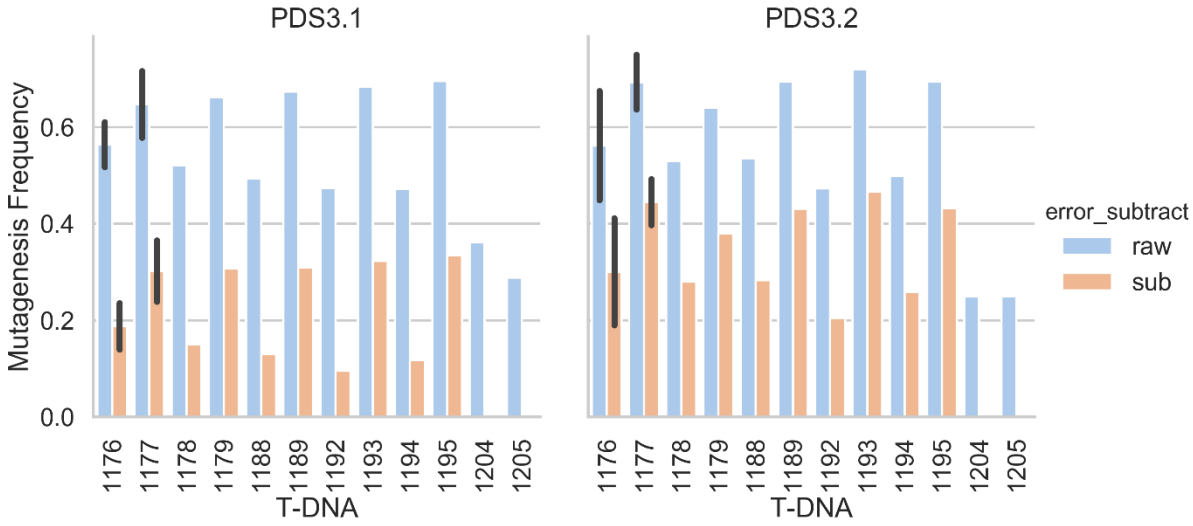


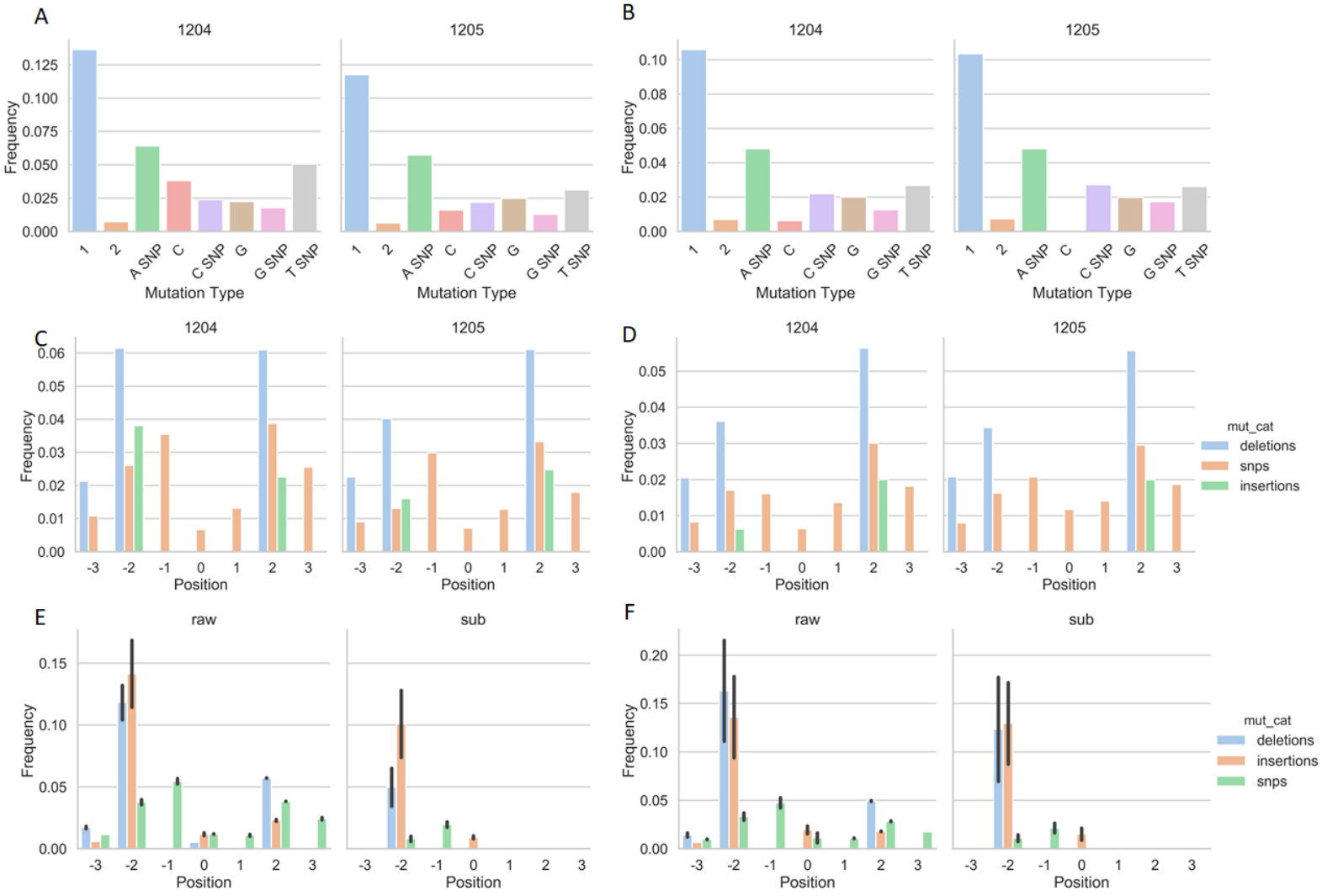
Figure 4-3 Conversion tract patterns from GT events grouped by directionality. ‘None’ indicates no conversion tracts present within that amplicon. Left Panel: GT SNP patterns collected at *PDS3.2* when delivered the divergent *PDS3.1* donor. Right Panel GT SNP patterns collected at *PDS3.1* when delivered the divergent *PDS3.2* donor.


A
B
C
D
E
F
G
H
Gene Targeting
Insertion
Deletion
SNPs
Frequency
Adjusted Frequency
Position
0
1000

Supplemental Figure 4-1 All mutations at all positions found in ONS amplicon reads. A-B) Two replicates of gRNA and donor treated sample. C-D) Two replicates of donor-only control sample. E-F) Samples as in A-B with mutations from D subtracted to show presence of nuclease-specific mutations. (G) Mutations in a donor-only control (C) after subtraction of mutations from a second donor-only control (D), demonstrating low level of background post-subtraction. (H) Donor-only control with all non-GT mutations from itself subtracted, leaving only GT events.



Supplemental Figure 4-2 Mutations at both targets for all treatments before and after error subtraction. Raw frequencies (blue) correspond to those without any processing and sub frequencies correspond the frequencies after subtraction on a per-mutation basis from sample 1204 (non-nuclease control). T-DNA details can be found in Supplemental Table 1.


A
1204
1205
Frequency
Mutation Type
1
2
A SNP
C
C SNP
G
G SNP
T SNP
B
C
D
Position
mut_cat
deletions
snps
insertions
E
raw
sub
F
insertions

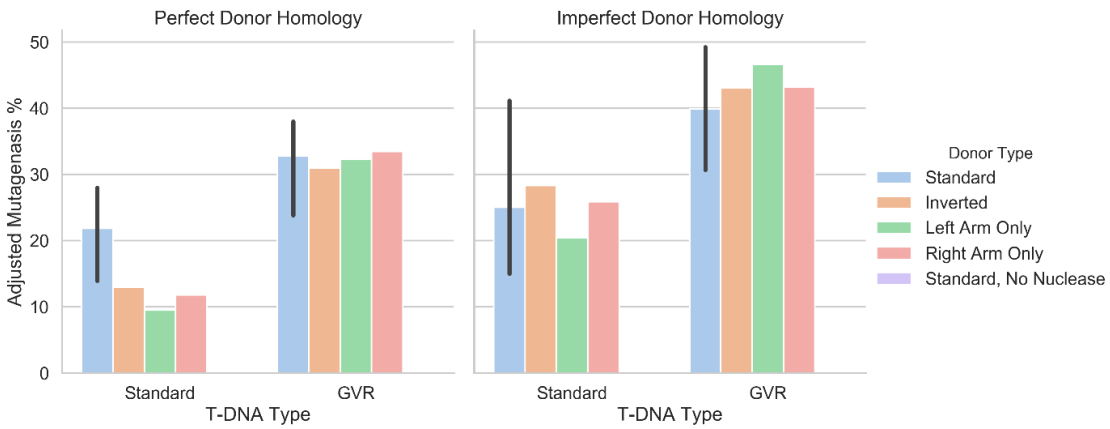
Supplemental Figure 4-3. Mutation profiles at target site gathered by Nanopore sequencing. (A) Frequency of individual mutations at target site above 0.5% at *PDS31* following treatment with donor alone control (1204) and donor alone plus GVR control (1205). Single numbers (1/2) indicate deletions of that size, letters indicate an insertion (C, (G), and letters followed by SNP indicate a SNP (C SNP, G SNP) at a position within the target region. (B) As in A, but at *PDS3.2*. (C) Mutations found at each specific position relative to the target site at *PDS3.1* with the same treatments found in A. Deletions (blue), SNPs (tan), and insertions (green) and their relative frequency at each positition. (D) As in C, but at *PDS3.2*. (E) As in C-D, but a nuclease + donor treatment. Left Panel: Frequencies prior to error subtraction (raw). Right Panel: frequencies post subtraction (sub). Threshold of 0.5% applied after error subtraction. (F) As in E, but at *PDS3.2*. Note presence of nuclease-specific mutations in both E and F.

PF \ LD	0	1	2	3	4	5	6	7	8	9	10	11	12	13	14	15	16	17	18	19	20
0	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.01	0.03	0.06	0.18	0.63	0.63	0.63	0.63
1	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.01	0.03	0.11	0.29	1.06				
2	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.01	0.04	0.17						
3	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.02	0.06	0.22						
4	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.01	0.02	0.08							
5	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.01	0.02	0.08							
6	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.01	0.03	0.11							
7	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.01	0.03	0.13							
8	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.01	0.04	0.16							
9	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.01	0.05	0.18							
10	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.02	0.06	0.21							
11	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.02	0.06								
12	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.02	0.08								
13	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.02	0.08								
14	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.03	0.10								
15	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.03	0.10								
16	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.03	0.10								
17	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.03	0.11								
18	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.03	0.12								
19	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.03	0.15								
20	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.03	0.15								

Supplemental Figure 4-4 Fuzz testing output for negative control. Frequency of GT (z-axis) found in donor-only treatment (1204) with given ‘fuzz’ parameters (LD: Levenshtein Distance, PF: Position Fuzz).

PF \ LD	0	1	2	3	4	5	6	7	8	9	10	11	12	13	14	15	16	17	18	19	20
0	0.02	0.03	0.03	0.04	0.04	0.04	0.04	0.04	0.04	0.04	0.04	0.04	0.05	0.06	0.12	0.32	0.76	4.67	4.67	4.67	4.67
1	0.02	0.03	0.04	0.04	0.04	0.04	0.05	0.05	0.05	0.05	0.05	0.05	0.07	0.11	0.30	0.73					
2	0.02	0.03	0.04	0.04	0.04	0.05	0.05	0.05	0.05	0.05	0.05	0.06	0.08	0.14	0.42						
3	0.02	0.03	0.04	0.04	0.04	0.05	0.05	0.05	0.05	0.06	0.06	0.07	0.09	0.17	0.51						
4	0.02	0.03	0.04	0.04	0.04	0.05	0.05	0.05	0.05	0.06	0.06	0.07	0.10	0.20	0.58						
5	0.02	0.03	0.04	0.04	0.04	0.05	0.05	0.05	0.05	0.06	0.06	0.07	0.11	0.22							
6	0.02	0.03	0.04	0.04	0.04	0.05	0.05	0.05	0.05	0.06	0.07	0.08	0.12	0.27							
7	0.02	0.03	0.04	0.04	0.04	0.05	0.05	0.05	0.05	0.06	0.07	0.08	0.13	0.29							
8	0.02	0.03	0.04	0.04	0.04	0.05	0.05	0.05	0.05	0.06	0.07	0.08	0.14	0.33							
9	0.02	0.03	0.04	0.04	0.04	0.05	0.05	0.05	0.05	0.06	0.07	0.08	0.15	0.35							
10	0.02	0.03	0.04	0.04	0.04	0.05	0.05	0.05	0.05	0.06	0.07	0.09	0.16	0.38							
11	0.02	0.03	0.04	0.04	0.04	0.05	0.05	0.05	0.05	0.06	0.07	0.09	0.17	0.40							
12	0.02	0.03	0.04	0.04	0.04	0.05	0.05	0.05	0.05	0.06	0.07	0.09	0.18	0.45							
13	0.02	0.03	0.04	0.04	0.04	0.05	0.05	0.05	0.05	0.06	0.07	0.09	0.18	0.47							
14	0.02	0.03	0.04	0.04	0.04	0.05	0.05	0.05	0.05	0.06	0.07	0.10	0.20	0.50							
15	0.02	0.03	0.04	0.04	0.04	0.05	0.05	0.05	0.05	0.06	0.07	0.10	0.20	0.52							
16	0.02	0.03	0.04	0.04	0.04	0.05	0.05	0.05	0.05	0.06	0.07	0.10	0.21	0.53							
17	0.02	0.03	0.04	0.04	0.04	0.05	0.05	0.05	0.05	0.06	0.07	0.10	0.22	0.56							
18	0.02	0.03	0.04	0.04	0.04	0.05	0.05	0.05	0.05	0.06	0.07	0.11	0.22								
19	0.02	0.03	0.04	0.04	0.04	0.05	0.05	0.05	0.05	0.06	0.07	0.11	0.25								
20	0.02	0.03	0.04	0.04	0.04	0.05	0.05	0.05	0.05	0.06	0.07	0.11	0.26								

Supplemental Figure 4-5 Fuzz testing output for gene targeting sample. Frequency of GT (z-axis) found in nuclease and donor treatment (1176) with given ‘fuzz’ parameters (LD: Levenshtein Distance, PF: Position Fuzz).

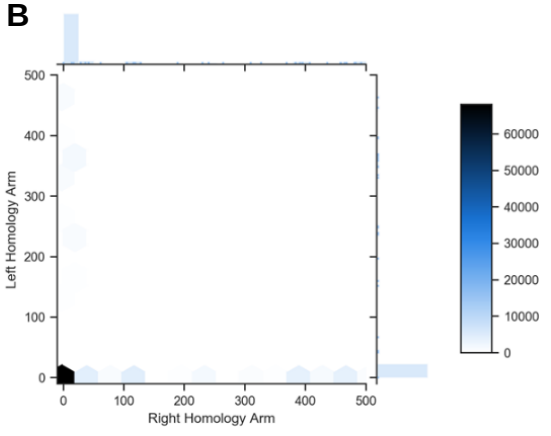


Supplemental Figure 4-6 Targeted mutagenesis frequency is consistent between treatments. Treatments with perfect donor homology (left panel) have consistent frequencies of targeted mutagenesis when T-DNA type (standard or GVR) is taken into consideration, regardless of the type of donor delivere. The same is true of imperfect donor homology treatments (right panel).

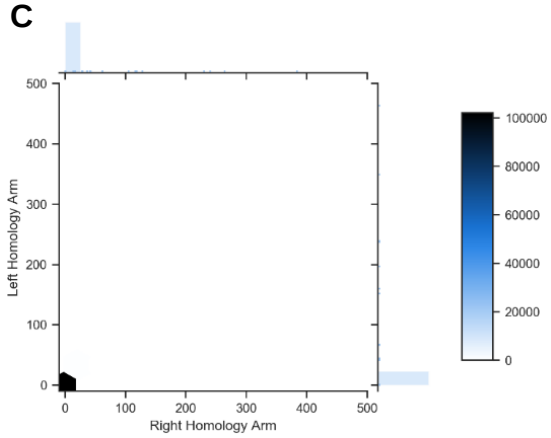
A

Pattern	Count
OOOOOOOOOOOOOOOOOOOOOOOOOOOOOOOOOOOOOOOOOOOOOOOOOOOOOOOOO	67526
OOOOOOOOOOOOOOOOOOOOOOOOOOOOOO1OOOOOOOOOOOOOOOOOOOOOOOOOO	3403
OOOOOOOOOOOOOOOOOOOOOOOOOOOOOOOO1OOOOOOOOOOOOOOOOOOOOOOOO	2321
OOOOOOOOOOOOOOOOOOOOOOOOOOOOOOOOOOOOOOOOOOOOOOOOOOOOOOOO	1872
OOOOOOOOOOOOOOOOOOOOOOOOOOOOOOOOOOOOOOOOOOOOOOOOOOOOOOO	1488
OOOOOOOOOOOOOOOOOOOOOOOOOOOOOOOOOOOOOOOOOOOOOOOOOOOOOOO	1180
OOOOOOOOOOOOOOOOOOOOOOOOOOOOOOOOOOOOOOOOOOOOOOOOOOOOOOO	1115
OOOOOOOOOOOOOOOOOOOOOOOOOOOOOOOOOOOOOOOOOOOOOOOOOOOOOOO	938
OOOOOOOOOOOOOOOOOOOOOOOOOOOOOOOOOOOOOOOOOOOOOOOOOOOOOOO	853
OOOOOOOOOOOOOOOOOOOOOOOOOOOOOOOOOOOOOOOOOOOOOOOOOOOOOOO	768
OOOOOOOOOOOOOOOOOOOOOOOOOOOOOOOOOOOOOOOOOOOOOOOOOOOOOOO	762
OOOOOOOOOOOOOOOOOOOOOOOOOOOOOOOOOOOOOOOOOOOOOOOOOOOOOOO	710
OOOOOOOOOOOOOOOOOOOOOOOOOOOOOOOOOOOOOOOOOOOOOOOOOOOOOOO	635
OOOOOOOOOOOOOOOOOOOOOOOOOOOOOOOOOOOOOOOOOOOOOOOOOOOOOOO	632
OOOOOOOOOOOOOOOOOOOOOOOOOOOOOOOOOOOOOOOOOOOOOOOOOOOOOOO	573
OOOOOOO1OOOOOOOOOOOOOOOOOOOOOOOOOOOOOOOOOOOOOOOOOOOOOOO	517
OOOOOOOO1OOOOOOOOOOOOOOOOOOOOOOOOOOOOOOOOOOOOOOOOOOOOOO	489
OOOOOOOOOOOOOOOOOOO1OOOOOOOOOOOOOOOOOOOOOOOOOOOOOOOOOOO	464
OOOOOOOOOO1OOOOOOOOOOOOOOOOOOOOOOOOOOOOOOOOOOOOOOOOOOOO	437
OOOOOOOOOOOOOOOOOOOOOOOO1OOOOOOOOOOOOOOOOOOOOOOOOOOOOOO	411

B



C

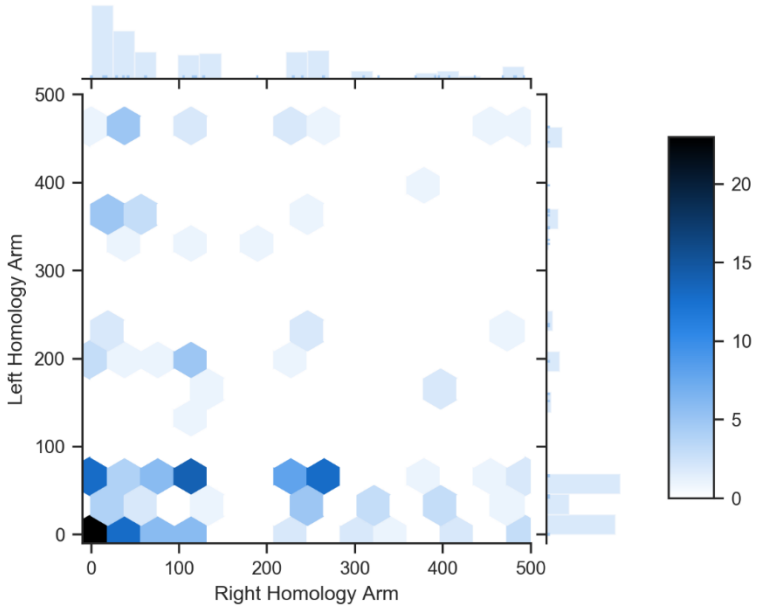


Supplemental Figure 4-7 Non-GT SNP patterns and are due to sequencing error and readily subtracted. (A) 20 most common SNP patterns at *PDS3.1* and their count in non-GT reads in a sample delivered gRNA and donor. (B) Conversion tracks from panel A visualized when using the outermost SNP to designate the conversion tracts end. (C) Conversion tracks from panel A visualized when ignoring SNPs that have three or more consecutive internal SNPs in order to reduce the impact of sequencing error on conversion tract analysis.

A

Pattern	Count
00000000000000000000000000000000000000000000000000000000000000	20
00000000000000000000011111000000000000000000000000000000000000	9
00000000000000000000000111111000000000000000000000000000000000	6
00000000000000000000011111111111111111000000000000000000000000	5
00000000000000000000000000010000000000000000000000000000000000	5
00000000000000000000000111100000000000000000000000000000000000	3
00000000000000000000000011000000000000000000000000000000000000	3
00000000010000000000011111111100000000000000000000000000000000	2
00000000000000000011101100000000000000000000000000000000000000	2
00000000000000000000111111111011110000000000000000000000000000	2
01000000000000000000000000010000000000000000000000000000000000	2
00000000000000000000000000110000000000000000000000000000000000	2
00000000000000000000000000010000000000000000000000000000000000	2
00000000000000000000011111111111000000000000000000000000000000	2
00000000000000000000000000010000000000000000000000000000000000	2
00000000000000000000111100000000000000000000000000000000000000	2
00000000000000000000111101100110000000000000000000000000000000	1
00000000000000000011111111010111100000000000000000000000000000	1
00000000000000000011111000100000000000000000000000000000000000	1
11111110111110000111111111011111111000000000000000000000000000	1

B



Supplemental Figure 4-8 GT-specific SNP patterns are consistent with known GT mechanisms. (A) 20 most common SNP patterns and their count in GT reads in a sample delivered gRNA and donor. (B) Conversion tracks from panel A visualized when using the outermost SNP to designate the conversion tracts end. Contrast with main figure 4-3B, which is derived from the same pattern dataset with a filter applied to ignore SNPs that contain 3 or more nonconsecutive SNPs.

T-DNA Used in Study	Description	Vectors/Amplicons used for Golden Gates to Create Vectors			
pPPA1176	PDS gRNA and PDS3.1 Donor	pTRANS_220d	pMOD_A0000d	pMM100	Cloning PCR Amplicon 1
pPPA1177	PDS gRNA and PDS3.1 Donor and GVR	pTRANS_221	pMOD_A0000d	pMM100	Cloning PCR Amplicon 1
pPPA1178	PDS gRNA and PDS3.2 Donor	pTRANS_220d	pMOD_A0000d	pMM100	Cloning PCR Amplicon 2
pPPA1179	PDS gRNA and PDS3.2 Donor and GVR	pTRANS_221	pMOD_A0000d	pMM100	Cloning PCR Amplicon 2
pPPA1188	PDS gRNA and Inverted PDS3.1 Donor	pTRANS_220d	pMOD_A0000d	pMM100	Cloning PCR Amplicon 3
pPPA1189	PDS gRNA and Inverted PDS3.1 Donor and GVR	pTRANS_221	pMOD_A0000d	pMM100	Cloning PCR Amplicon 3
pPPA1192	PDS gRNA and Left Arm Only PDS3.1 Donor	pTRANS_220d	pMOD_A0000d	pMM100	Cloning PCR Amplicon 4
pPPA1193	PDS gRNA and Left Arm Only PDS3.1 Donor and GVR	pTRANS_221	pMOD_A0000d	pMM100	Cloning PCR Amplicon 4
pPPA1194	PDS gRNA and Right Arm Only PDS3.1 Donor	pTRANS_220d	pMOD_A0000d	pMM100	Cloning PCR Amplicon 5
pPPA1195	PDS gRNA and Right Arm Only PDS3.1 Donor and GVR	pTRANS_221	pMOD_A0000d	pMM100	Cloning PCR Amplicon 5
pPPA1204	PDS3.1 Donor	pTRANS_220d	pMOD_A0000d	pMOD_B0000d	Cloning PCR Amplicon 1
pPPA1205	PDS3.1 Donor and GVR	pTRANS_221	pMOD_A0000d	pMOD_B0000d	Cloning PCR Amplicon 1

Supplemental Table 4-1. T-DNAs used in the study, descriptions, and the components from which they were created.

Amplicon	Template	Primer 1	Primer 2
Cloning PCR Amplicon 1	Synthesized *PDS3.1* Fragment	GCGCCCACCTGCAACGCCGGGATTTTGGTTTCTTTGGTTA	TTCATCACCTGCTAGTCACTCAAACTAGTTTTCAAACCGC
Cloning PCR Amplicon 2	Synthesized *PDS3.2* Fragment	GCGCCCACCTGCAACGCCGGATTTTGGTTTCTTTGGTTAC	TTCATCACCTGCTAGTCACTAATAGCTCAAAACAAACTAA
Cloning PCR Amplicon 3	Synthesized *PDS3.1* Fragment	TTCATCACCTGCTAGTCACTGATTTTGGTTTCTTTGGTTA	GCGCCCACCTGCAACGCCGGCAAACTAGTTTTCAAACCGC
Cloning PCR Amplicon 4	Synthesized *PDS3.1* Fragment	GCGCCCACCTGCAACGCCGGGATTTTGGTTTCTTTGGTTA	TTCATCACCTGCTAGTCACTATTGGACAGACCATGGATGG
Cloning PCR Amplicon 5	Synthesized *PDS3.1* Fragment	GCGCCCACCTGCAACGCCGGATCCATGGTCTGTCCAATAT	TTCATCACCTGCTAGTCACTCAAACTAGTTTTCAAACCGC

Supplemental Table 4-2. Cloning PCR Amplicons and their templates and primers.

Name	Sequence
Benthi PDS3.1 F-1	AAGAAAGTTGTCGGTGTCTTTGTG **AATGGTGGGACATTTTTGGA**
Benthi PDS3.1 F-2	TCGATTCCGTTTGTAGTCGTCTGT **AATGGTGGGACATTTTTGGA**
Benthi PDS3.1 F-3	GAGTCTTGTGTCCCAGTTACCAGG **AATGGTGGGACATTTTTGGA**
Benthi PDS3.1 F-4	TTCGGATTCTATCGTGTTTCCCTA **AATGGTGGGACATTTTTGGA**
Benthi PDS3.1 F-5	CTTGTCCAGGGTTTGTGTAACCTT **AATGGTGGGACATTTTTGGA**
Benthi PDS3.1 F-6	TTCTCGCAAAGGCAGAAAGTAGTC **GCTAGCTTATGAGGTGAAGC**
Benthi PDS3.1 R-1	GTGTTACCGTGGGAATGAATCCTT **GCTAGCTTATGAGGTGAAGC**
Benthi PDS3.1 R-2	TTCAGGGAACAAACCAAGTTACGT **GCTAGCTTATGAGGTGAAGC**
Benthi PDS3.1 R-3	AACTAGGCACAGCGAGTCTTGGTT **GCTAGCTTATGAGGTGAAGC**
Benthi PDS3.1 R-4	AAGCGTTGAAACCTTTGTCCTCTC **GCTAGCTTATGAGGTGAAGC**
Benthi PDS3.1 R-5	GTTTCATCTATCGGAGGGAATGGA **GCTAGCTTATGAGGTGAAGC**
Benthi PDS3.1 R-6	CAGGTAGAAAGAAGCAGAATCGGA **GCTAGCTTATGAGGTGAAGC**
Benthi PDS3.2 F-1	AAGAAAGTTGTCGGTGTCTTTGTG **TAGCAAACAAATGACCCACC**
Benthi PDS3.2 F-2	TCGATTCCGTTTGTAGTCGTCTGT **TAGCAAACAAATGACCCACC**
Benthi PDS3.2 F-3	GAGTCTTGTGTCCCAGTTACCAGG **TAGCAAACAAATGACCCACC**
Benthi PDS3.2 F-4	TTCGGATTCTATCGTGTTTCCCTA **TAGCAAACAAATGACCCACC**
Benthi PDS3.2 F-5	CTTGTCCAGGGTTTGTGTAACCTT **TAGCAAACAAATGACCCACC**
Benthi PDS3.2 F-6	TTCTCGCAAAGGCAGAAAGTAGTC **TAGCAAACAAATGACCCACC**
Benthi PDS3.2 R-1	GTGTTACCGTGGGAATGAATCCTT **TGTGGGACAACTCAACACCC**
Benthi PDS3.2 R-2	TTCAGGGAACAAACCAAGTTACGT **TGTGGGACAACTCAACACCC**
Benthi PDS3.2 R-3	AACTAGGCACAGCGAGTCTTGGTT **TGTGGGACAACTCAACACCC**
Benthi PDS3.2 R-4	AAGCGTTGAAACCTTTGTCCTCTC **TGTGGGACAACTCAACACCC**
Benthi PDS3.2 R-5	GTTTCATCTATCGGAGGGAATGGA **TGTGGGACAACTCAACACCC**
Benthi PDS3.2 R-6	CAGGTAGAAAGAAGCAGAATCGGA **TGTGGGACAACTCAACACCC**

Supplemental Table 4-3. Barcoded oligos used for PDS3.1 and PDS3.2 to enable multiplexing up to 36 samples for each target. Names indicate the target (*PDS3.1* or *PDS3.2*), the directionality (Forward or Reverse), and which barcode (1-6). Each forward and reverse primer for each target has a unique barcode but the barcodes are repeated between *PDS3.1* and *PDS3.2* oligo sets.

Supplemental Sequence 4-1. Synthesized *PDS3.1* Donor

GATTTTGGTTTCTTTGGTTACATCAGCTGAATGCTTTGCTTGAGAAAAGCTCT
CTTTTTCCCGTTTAGGATCTTGTTTATTTGCTTTCGTTTTTCTACTCGTTTGAA
TTTTAACTTGATTTTGTGGGTGAAGGCTAATTTTTCTCATAGTGTAAGAACAA
GTTTCATATGTACTGTAAAAGCTAGAATCTTTTTTACTTTTGCATATAAATTTG
TGTAATAAATGCTTAAGAACCAGAATATTTGAAAAAGATAAGGAATTTTGCAT
AGTATTTAGGTTCACAAGTGGGACAATCTTCTTACACTGAAATATCTTTATGT
CAGGCTTAATTTACTGCTATCTTGTTCAATAAAATGCCCCAAATTGGACTTGT
TTCTGCCGTTAATTTGAGAGTCCAAGGTAATTCAGCTTATCTTTGGAGCTCG
AGGTCTTCGTTGGGAACTGAAAGTCAAGATGTTTGCTTGCAAAGGAATTTGT
TATGTTTTGGTAGTAGCGACTCCATCCATGGTCTGTCCAATATGGGGCATAA
GTTAAGGATTCGTACTCCAAGTGCCACGACCCGAAGATTGACAAAGGACTT
TAATCCTTTAAAGGTTTGTTTTGAATGCGAAAGTGTGATGCTGGATTTATGAT
CGTGGGCATATATCCTCTAAAATAAGAGATGTATATCTTGCCATTCAGGTAG
TCTGCATTGATTATCCAAGACCAGAGCTAGACAATACAGTTAACTATTTGGA
GGCGGCGTTATTATCATCATCGTTTCGTACTTCCTCACGCCCAACTAAACCA
TTGGAGATTGTTATTGCTGGTGCAGGTGATTTTTTCCAGCCATCTATATTTGT
AGTTTTCATTTTTCTTTCTTTGGAAGGAAGATCATTCTATTAGTTATATTATCA
CTAGAATATTTACCTGTACATTCTTTTCTGATTAACTGTTTTGGACCGCAAAA
TTTTAGGTTCTTACTTCTTCGCCATTTTGCAACTAATCAGCAATTAGGAGCGG
TTTGAAAACTAGTTTG

Supplemental Sequence 4-2. Synthesized *PDS3.2* Donor

ATTTTGGTTTCTTTGGTTACATCAGCTGAATGCTTTACTTGAGAAAAGCTTTC
TCCTTTTCCCGTTTAGGATCTTGTTTATTTGCTTTCGTTTTTCTACTCGTTAAA
ATTTTAACTTGATTTTGTGGGTGAATTATAACTTTACTCATAGTGCGAGAACA
AGTTTCGTATGGACTGTAAAAGCTAGAATCTTTTTTACTTTTGCATATAAATTT
GTGTAATAAATGCTTAAGAACCAGAATATTGAAAAAACAAAGGAATTCTACAT
AGTATTTAGGTTCACAAGTGGGACAATCTTCTTACAGTGAAATATCTTTATGT
CAGGCTTAATTTACTGCTATTTTGTTCAGTAAAATGCCCCAAATTGGACTTGT
TTCTGCCGTTAATTTGAGAGTCCAAGGTAATTCAGCTTATCTTTGGAGCTCG
AGGTCTTCTTTGGGAACTGAAAGTCAAGATGGTCGCTTGCAAAGGAATTTGT
TATGTTTTGGTAGTAGCGACTCCATCCATGGTCTGTCCAATATGGGGCATAA
GTTTAGAATTCGTACTCCCAGTGCCATGACCAGAAGATTGACAAAGGACTTC
AATCCTTTAAAGGTTTGTTTTGAATGCGAAAGTGTGATGCTGAATTTATGATC
ACGAGCATATATTCTCTAAAATAAGATATCTTGCCATTCAGGTAGTCTGCATT
GATTATCCAAGACCGGAGCTAGACAATACAGTTAACTATTTGGAGGCGGCG
TTATCATCATCATCATTTCGTACTTCCTCACGCCCAACAAAACCATTGGAGAT
TGTTATTGCTGGTGCAGGTGATTTTTTCCAGTCATCTATATTTGTAGTCTTCA
TTTTTCTTTCTTTGGAAGGAAGATCATTCTATTAGTTGTATTATCACTAGAACA
TTTATTGTGCATTCTTTTCTTATTAACTGTTTTGGACCGCAAAATTTTAAGTTC
TTACTTCTTCGCCTCCCAACTGATTAGATTAGGAGTGATTTGAAAATTAGTTT
GTTTTGAGCTATT

CHAPTER FIVE

Conclusions and Future Directions

Conclusions

The study of plant gene targeting has long been fragmented by technical idiosyncrasies between species and research groups. Distinct combinations of techniques prevent clear communication of results, resulting in numerous technical islands. Here I developed a novel approach using Oxford Nanopore Sequencing (ONS) to quantify gene targeting at any conceivable genomic target. The low cost, rapid turnaround, and high throughput of this approach make it ideal for quantification of plant genome editing outcomes across species and technical barriers. This required the creation of a bioinformatic pipeline capable of quantifying genome editing outcomes and overcoming the principle shortcoming of ONS – sequencing error (Chapter 3). Using this pipeline, I observed the significant impact of geminiviral replicons and imperfect donor homology arms on gene targeting frequencies while collecting hundreds of conversion tracts from gene targeting events, garnering mechanistic insights (Chapter 4). This work establishes a framework from which gene targeting in any organism may be robustly dissected at the molecular level at an unprecedented pace.

Future Directions

Establishing an Oxford Nanopore-based gene targeting (GT) analysis pipeline removes several impediments to the optimization of plant gene targeting. This will facilitate the analysis of key parameters on gene targeting frequency, particularly the type DNA damage used to initiate GT and donor molecule properties. Nuclease cut types may significantly alter DNA repair outcomes, therefore Cas9 (blunt cutting), Cas12a (creates overhangs through staggered cuts), and ssDNA nicks using both systems should be compared. Examining basic properties of the donor molecule, such as the ideal donor arm size and the effect of insertion size on efficiency, are also vital experiments. Beyond reagent

variants, the epigenetic context of target sites should be examined, specifically comparing targets in repressed heterochromatic regions, constitutively activated regions, and regions with variable or tissue-specific expression patterns. Finally, the addition of a graphic user interface to PANGEA will drastically increase ease of use and allow for broad adoption. Together, these insights into reagent and target parameters paired with readily usable analysis software will drastically reduce the entry barrier into precision plant genome editing and result in widespread protocol improvements.

References

Acinas, Silvia G., Ramahi Sarma-Rupavtarm, Vanja Klepac-Ceraj, and Martin F. Polz. 2005. “PCR-Induced Sequence Artifacts and Bias: Insights from Comparison of Two 16S RRNA Clone Libraries Constructed from the Same Sample.” *Applied and Environmental Microbiology* 71 (12): 8966–69. https://doi.org/10.1128/AEM.71.12.8966-8969.2005.

Altpeter, Fredy, Nathan M. Springer, Laura E. Bartley, Ann E. Blechl, Thomas P. Brutnell, Vitaly Citovsky, Liza J. Conrad, et al. 2016. “Advancing Crop Transformation in the Era of Genome Editing[OPEN].” *The Plant Cell* 28 (7): 1510–20. https://doi.org/10.1105/tpc.16.00196.

Ayar, Ayhan, Sophie Wehrkamp-Richter, Jean-Baptiste Laffaire, Samuel Le Goff, Julien Levy, Sandrine Chaignon, Hajer Salmi, et al. 2013. “Gene Targeting in Maize by Somatic Ectopic Recombination.” *Plant Biotechnology Journal* 11 (3): 305–14. https://doi.org/10.1111/pbi.12014.

Baltes, Nicholas J., Javier Gil-Humanes, Tomas Cermak, Paul A. Atkins, and Daniel F. Voytas. 2014. “DNA Replicons for Plant Genome Engineering.” *The Plant Cell* 26 (1): 151–63. https://doi.org/10.1105/tpc.113.119792.

Beetham, P. R., P. B. Kipp, X. L. Sawycky, C. J. Arntzen, and G. D. May. 1999. “A Tool for Functional Plant Genomics: Chimeric RNA/DNA Oligonucleotides Cause in Vivo Gene-Specific Mutations.” *Proceedings of the National Academy of Sciences* 96 (15): 8774–78. https://doi.org/10.1073/pnas.96.15.8774.

Brinkman, Eva Karina, and Bas van Steensel. 2019. “Rapid Quantitative Evaluation of CRISPR Genome Editing by TIDE and TIDER.” In *CRISPR Gene Editing: Methods and Protocols*, edited by Yonglun Luo, 29–44. Methods in Molecular Biology. New York, NY: Springer. https://doi.org/10.1007/978-1-4939-9170-9_3.

Butler, Nathaniel M., Nicholas J. Baltes, Daniel F. Voytas, and David S. Douches. 2016. “Geminivirus-Mediated Genome Editing in Potato (Solanum Tuberosum L.) Using Sequence-Specific Nucleases.” *Frontiers in Plant Science* 7. https://doi.org/10.3389/fpls.2016.01045.

Canaj, Hera, Jeffrey A. Hussmann, Han Li, Kyle A. Beckman, Leeanne Goodrich, Nathan H. Cho, Yucheng J. Li, et al. 2019. “Deep Profiling Reveals Substantial Heterogeneity of Integration Outcomes in CRISPR Knock-in Experiments.” Preprint. Genomics. https://doi.org/10.1101/841098.

Čermák, T., S. J. Curtin, J. Gil-Humanes, R. Čegan, T. J. Y. Kono, E. Konečná, J. J. Belanto, et al. 2017. “A Multipurpose Toolkit to Enable Advanced Genome Engineering in Plants.” *The Plant Cell* 29 (6): 1196–1217. https://doi.org/10.1105/tpc.16.00922.

Čermák, Tomáš, Nicholas J. Baltes, Radim Čegan, Yong Zhang, and Daniel F. Voytas. 2015. “High-Frequency, Precise Modification of the Tomato Genome.” *Genome Biology* 16 (1): 232. https://doi.org/10.1186/s13059-015-0796-9.

Chandrasegaran, Srinivasan, and Dana Carroll. 2016. “Origins of Programmable Nucleases for Genome Engineering.” *Journal of Molecular Biology*, Engineering Tools and Prospects in Synthetic Biology, 428 (5, Part B): 963–89. https://doi.org/10.1016/j.jmb.2015.10.014.

Cho, Seung Woo, Sojung Kim, Jong Min Kim, and Jin-Soo Kim. 2013. “Targeted Genome Engineering in Human Cells with the Cas9 RNA-Guided Endonuclease.” *Nature Biotechnology* 31 (3): 230–32. https://doi.org/10.1038/nbt.2507.

Cody, Jon P., Nathaniel D. Graham, and James A. Birchler. 2017. “BiBAC Modification and Stable Transfer into Maize (Zea Mays) Hi-II Immature Embryos via Agrobacterium-Mediated Transformation.” *Current Protocols in Plant Biology* 2 (4): 350–69. https://doi.org/10.1002/cppb.20061.

Cong, L., F. A. Ran, D. Cox, S. Lin, R. Barretto, N. Habib, P. D. Hsu, et al. 2013. “Multiplex Genome Engineering Using CRISPR/Cas Systems.” *Science* 339 (6121): 819–23. https://doi.org/10.1126/science.1231143.

Day, Elizabeth, Paul H. Dear, and Frank McCaughan. 2013. “Digital PCR Strategies in the Development and Analysis of Molecular Biomarkers for Personalized Medicine.” *Methods*, Transcriptional Biomarkers, 59 (1): 101–7. https://doi.org/10.1016/j.ymeth.2012.08.001.

Decaestecker, Ward, Rafael Andrade Buono, Marie Pfeiffer, Nick Vangheluwe, Joris Jourquin, Mansour Karimi, Gert van Isterdael, Tom Beeckman, Moritz K. Nowack, and Thomas B. Jacobs. 2019. “CRISPR-TSKO: A Technique for Efficient Mutagenesis in Specific Cell Types, Tissues, or Organs in Arabidopsis.” *The Plant Cell*, January, tpc.00454.2019. https://doi.org/10.1105/tpc.19.00454.

Durr, Julius, Ranjith Papareddy, Keiji Nakajima, and Jose Gutierrez-Marcos. 2018. “Highly Efficient Heritable Targeted Deletions of Gene Clusters and Non-Coding Regulatory Regions in Arabidopsis Using CRISPR/Cas9.” *Scientific Reports* 8 (1): 4443. https://doi.org/10.1038/s41598-018-22667-1.

Eloe-Fadrosh, Emiley A., Natalia N. Ivanova, Tanja Woyke, and Nikos C. Kyrpides. 2016. “Metagenomics Uncovers Gaps in Amplicon-Based Detection of Microbial Diversity.” *Nature Microbiology* 1 (February): 15032. https://doi.org/10.1038/nmicrobiol.2015.32.

Emmanuel, Eyal, Elizabeth Yehuda, Cathy Melamed-Bessudo, Naomi Avivi-Ragolsky, and Avraham A Levy. 2006. “The Role of AtMSH2 in Homologous Recombination in Arabidopsis Thaliana.” *EMBO Reports* 7 (1): 100–105. https://doi.org/10.1038/sj.embor.7400577.

Emmanuel, Eyal, Elizabeth Yehuda, Cathy Melamed-Bessudo, Naomi Avivi-Ragolsky, and Avraham A. Levy. 2006. “The Role of AtMSH2 in Homologous Recombination in Arabidopsis Thaliana.” *EMBO Reports* 7 (1): 100–105. https://doi.org/10.1038/sj.embor.7400577.

Endo, Masaki, Masafumi Mikami, and Seiichi Toki. 2016. “Biallelic Gene Targeting in Rice.” *Plant Physiology* 170 (2): 667–77. https://doi.org/10.1104/pp.15.01663.

Fauser, Friedrich, Nadine Roth, Michael Pacher, Gabriele Ilg, Rocío Sánchez-Fernández, Christian Biesgen, and Holger Puchta. 2012. “In Planta Gene Targeting.” *Proceedings of the National Academy of Sciences* 109 (19): 7535–40. https://doi.org/10.1073/pnas.1202191109.

Gasiunas, Giedrius, Rodolphe Barrangou, Philippe Horvath, and Virginijus Siksnys. 2012. “Cas9–CrRNA Ribonucleoprotein Complex Mediates Specific DNA Cleavage for Adaptive Immunity in Bacteria.” *Proceedings of the National Academy of Sciences* 109 (39): E2579–86. https://doi.org/10.1073/pnas.1208507109.

Gaudelli, Nicole M., Alexis C. Komor, Holly A. Rees, Michael S. Packer, Ahmed H. Badran, David I. Bryson, and David R. Liu. 2017. “Programmable Base Editing of A•T to G•C in Genomic DNA without DNA Cleavage.” *Nature* 551 (7681): 464–71. https://doi.org/10.1038/nature24644.

Gil-Humanes, Javier, Yanpeng Wang, Zhen Liang, Qiwei Shan, Carmen V. Ozuna, Susana Sánchez-León, Nicholas J. Baltes, et al. 2017. “High-Efficiency Gene Targeting in Hexaploid Wheat Using DNA Replicons and CRISPR/Cas9.” *The Plant Journal* 89 (6): 1251–62. https://doi.org/10.1111/tpj.13446.

Gonzalez, Valentina, and Claudia P. Spampinato. 2020. “The Mismatch Repair Protein MSH6 Regulates Somatic Recombination in Arabidopsis Thaliana.” *DNA Repair* 87 (March): 102789. https://doi.org/10.1016/j.dnarep.2020.102789.

Hahn, Florian, Marion Eisenhut, Otho Mantegazza, and Andreas P. M. Weber. 2018. “Homology-Directed Repair of a Defective Glabrous Gene in Arabidopsis With Cas9-Based Gene Targeting.” *Frontiers in Plant Science* 9. https://doi.org/10.3389/fpls.2018.00424.

Hahn, Florian, Andrey Korolev, Laura Sanjurjo Loures, and Vladimir Nekrasov. 2019. “A Modular Cloning Toolkit for Genome Editing in Plants.” *BioRxiv*, August, 738021. https://doi.org/10.1101/738021.

Hanin, Moez, Sandra Volrath, Augustyn Bogucki, Markus Briker, Eric Ward, and Jerzy Paszkowski. 2001. “Gene Targeting in Arabidopsis.” *The Plant Journal* 28 (6): 671–77. https://doi.org/10.1046/j.1365-313x.2001.01183.x.

Harwood, Wendy, Nathalia Volpi e Silva, and Nicola J. Patron. 2017. “CRISPR-Based Tools for Plant Genome Engineering.” *Emerging Topics in Life Sciences* 1 (2): 135–49. https://doi.org/10.1042/ETLS20170011.

Hendel, Ayal, Eric J. Kildebeck, Eli J. Fine, Joseph T. Clark, Niraj Punjya, Vittorio Sebastiano, Gang Bao, and Matthew H. Porteus. 2014. “Quantifying Genome-Editing Outcomes at Endogenous Loci with SMRT Sequencing.” *Cell Reports* 7 (1): 293–305. https://doi.org/10.1016/j.celrep.2014.02.040.

Heyer, Wolf-Dietrich, Kirk T. Ehmsen, and Jie Liu. 2010. “Regulation of Homologous Recombination in Eukaryotes.” *Annual Review of Genetics* 44 (1): 113–39. https://doi.org/10.1146/annurev-genet-051710-150955.

Hohn, Barbara, and Holger Puchta. 1999. “Gene Therapy in Plants.” *Proceedings of the National Academy of Sciences of the United States of America* 96 (15): 8321–23.

Huang, Teng-Kuei. 2019. “CRISPR/Cas-Mediated Gene Targeting in Plants: Finally a Turn for the Better for Homologous Recombination.” *Plant Cell Reports*, 11.

Huang, Teng-Kuei, and Holger Puchta. 2019. “CRISPR/Cas-Mediated Gene Targeting in Plants: Finally a Turn for the Better for Homologous Recombination.” *Plant Cell Reports* 38 (4): 443–53. https://doi.org/10.1007/s00299-019-02379-0.

Inagaki, Soichi, Takamasa Suzuki, Masa-aki Ohto, Hiroko Urawa, Takashi Horiuchi, Kenzo Nakamura, and Atsushi Morikami. 2006. “Arabidopsis TEBICHI, with Helicase and DNA Polymerase Domains, Is Required for Regulated Cell Division and Differentiation in Meristems.” *The Plant Cell* 18 (4): 879–92. https://doi.org/10.1105/tpc.105.036798.

Jain, Miten, Sergey Koren, Karen H Miga, Josh Quick, Arthur C Rand, Thomas A Sasani, John R Tyson, et al. 2018. “Nanopore Sequencing and Assembly of a Human Genome with Ultra-Long Reads.” *Nature Biotechnology* 36 (4): 338–45. https://doi.org/10.1038/nbt.4060.

Jinek, M., K. Chylinski, I. Fonfara, M. Hauer, J. A. Doudna, and E. Charpentier. 2012. “A Programmable Dual-RNA-Guided DNA Endonuclease in Adaptive Bacterial Immunity.” *Science* 337 (6096): 816–21. https://doi.org/10.1126/science.1225829.

Kaeppler, Shawn M., Heidi F. Kaeppler, and Yong Rhee. 2000. “Epigenetic Aspects of Somaclonal Variation in Plants.” In *Plant Gene Silencing*, edited by M. A. Matzke and A. J. M. Matzke, 59–68. Dordrecht: Springer Netherlands. https://doi.org/10.1007/978-94-011-4183-3_4.

Kalle, Elena, Alexander Gulevich, and Christopher Rensing. 2013. “External and Semi-Internal Controls for PCR Amplification of Homologous Sequences in Mixed Templates.” *Journal of Microbiological Methods* 95 (2): 285–94. https://doi.org/10.1016/j.mimet.2013.09.014.

Kan, Yinan, and Eric A. Hendrickson. 2019. “Conversion Tract Analysis of Homology-Directed Genome Editing Using Oligonucleotide Donors.” In *DNA Repair: Methods and Protocols*, edited by Lata Balakrishnan and Jason A. Stewart, 131–44. Methods in Molecular Biology. New York, NY: Springer New York. https://doi.org/10.1007/978-1-4939-9500-4_7.

Kan, Yinan, Brian Ruis, Sherry Lin, and Eric A. Hendrickson. 2014. “The Mechanism of Gene Targeting in Human Somatic Cells.” *PLOS Genetics* 10 (4): e1004251. https://doi.org/10.1371/journal.pgen.1004251.

Kang, Beum-Chang, Jae-Young Yun, Sang-Tae Kim, YouJin Shin, Jahee Ryu, Minkyung Choi, Je Wook Woo, and Jin-Soo Kim. 2018. “Precision Genome Engineering through Adenine Base Editing in Plants.” *Nature Plants* 4 (7): 427–31. https://doi.org/10.1038/s41477-018-0178-x.

Karst, Søren M., Ryan M. Ziels, Rasmus H. Kirkegaard, Emil A. Sørensen, Daniel McDonald, Qiyun Zhu, Rob Knight, and Mads Albertsen. 2020. “Enabling High-Accuracy Long-Read Amplicon Sequences Using Unique Molecular Identifiers with Nanopore or PacBio Sequencing.” *BioRxiv*, January, 645903. https://doi.org/10.1101/645903.

Kim, Sang-Ic, and Stanton B. Gelvin. 2007. “Genome-Wide Analysis of Agrobacterium T-DNA Integration Sites in the Arabidopsis Genome Generated under Non-Selective Conditions.” *The Plant Journal* 51 (5): 779–91. https://doi.org/10.1111/j.1365-313X.2007.03183.x.

Kivioja, Teemu, Anna Vähärautio, Kasper Karlsson, Martin Bonke, Martin Enge, Sten Linnarsson, and Jussi Taipale. 2012. “Counting Absolute Numbers of Molecules Using Unique Molecular Identifiers.” *Nature Methods* 9 (1): 72–74. https://doi.org/10.1038/nmeth.1778.

Kluesner, Mitchell G., Derek A. Nedveck, Walker S. Lahr, John R. Garbe, Juan E. Abrahante, Beau R. Webber, and Branden S. Moriarity. 2018. “EditR: A Method to Quantify Base Editing from Sanger Sequencing.” *The CRISPR Journal* 1 (3): 239–50. https://doi.org/10.1089/crispr.2018.0014.

Koller, B H, and O Smithies. 1992. “Altering Genes in Animals by Gene Targeting.” *Annual Review of Immunology* 10 (1): 705–30. https://doi.org/10.1146/annurev.iy.10.040192.003421.

Komor, Alexis C., Yongjoo B. Kim, Michael S. Packer, John A. Zuris, and David R. Liu. 2016. “Programmable Editing of a Target Base in Genomic DNA without Double-Stranded DNA Cleavage.” *Nature* 533 (7603): 420–24. https://doi.org/10.1038/nature17946.

Kregten, Maartje van, Sylvia de Pater, Ron Romeijn, Robin van Schendel, Paul J. J. Hooykaas, and Marcel Tijsterman. 2016. “T-DNA Integration in Plants Results from Polymerase- θ -Mediated DNA Repair.” *Nature Plants* 2 (11): 1–6. https://doi.org/10.1038/nplants.2016.164.

Krehenwinkel, Henrik, Aaron Pomerantz, James B. Henderson, Susan R. Kennedy, Jun Ying Lim, Varun Swamy, Juan Diego Shoobridge, et al. 2019. “Nanopore Sequencing of Long Ribosomal DNA Amplicons Enables Portable and Simple Biodiversity Assessments with High Phylogenetic Resolution across Broad Taxonomic Scale.” *GigaScience* 8 (5). https://doi.org/10.1093/gigascience/giz006.

Kumar, Sandeep, Andrew Worden, Stephen Novak, Ryan Lee, and Joseph F. Petolino. 2016. “A Trait Stacking System via Intra-Genomic Homologous Recombination.” *Planta* 244 (5): 1157–66. https://doi.org/10.1007/s00425-016-2595-2.

Lacroix, Benoît, and Vitaly Citovsky. 2019. “Pathways of DNA Transfer to Plants from *Agrobacterium Tumefaciens* and Related Bacterial Species.” *Annual Review of Phytopathology* 57 (1): 231–51. https://doi.org/10.1146/annurev-phyto-082718-100101.

Laver, T., J. Harrison, P. A. O’Neill, K. Moore, A. Farbos, K. Paszkiewicz, and D. J. Studholme. 2015. “Assessing the Performance of the Oxford Nanopore Technologies MinION.” *Biomolecular Detection and Quantification* 3 (March): 1–8. https://doi.org/10.1016/j.bdq.2015.02.001.

Lemmon, Zachary H., Nathan T. Reem, Justin Dalrymple, Sebastian Soyk, Kerry E. Swartwood, Daniel Rodriguez-Leal, Joyce Van Eck, and Zachary B. Lippman. 2018. “Rapid Improvement of Domestication Traits in an Orphan

Crop by Genome Editing." *Nature Plants* 4 (10): 766–70. https://doi.org/10.1038/s41477-018-0259-x.
Li, Chao, Yuan Zong, Yanpeng Wang, Shuai Jin, Dingbo Zhang, Qianna Song, Rui Zhang, and Caixia Gao. 2018. "Expanded Base Editing in Rice and Wheat Using a Cas9-Adenosine Deaminase Fusion." *Genome Biology* 19 (May). https://doi.org/10.1186/s13059-018-1443-z.
Li, Heng. 2018. "Minimap2: Pairwise Alignment for Nucleotide Sequences." Edited by Inanc Birol. *Bioinformatics* 34 (18): 3094–3100. https://doi.org/10.1093/bioinformatics/bty191.
Li, Jun, Xiangbing Meng, Yuan Zong, Kunling Chen, Huawei Zhang, Jinxing Liu, Jiayang Li, and Caixia Gao. 2016. "Gene Replacements and Insertions in Rice by Intron Targeting Using CRISPR-Cas9." *Nature Plants* 2: 16139. https://doi.org/10.1038/nplants.2016.139.
Li, Liangliang, Martine Jean, and François Belzile. 2006. "The Impact of Sequence Divergence and DNA Mismatch Repair on Homeologous Recombination in Arabidopsis." *The Plant Journal* 45 (6): 908–16. https://doi.org/10.1111/j.1365-313X.2006.02657.x.
Li, Tingdong, Xinping Yang, Yuan Yu, Xiaomin Si, Xiawan Zhai, Huawei Zhang, Wenxia Dong, Caixia Gao, and Cao Xu. 2018. "Domestication of Wild Tomato Is Accelerated by Genome Editing." *Nature Biotechnology* 36 (12): 1160–63. https://doi.org/10.1038/nbt.4273.
Lieberman-Lazarovich, Michal, Cathy Melamed-Bessudo, Sylvia de Pater, and Avraham A. Levy. 2013. "Epigenetic Alterations at Genomic Loci Modified by Gene Targeting in Arabidopsis Thaliana." *PLoS ONE* 8 (12). https://doi.org/10.1371/journal.pone.0085383.
Lisby, Michael, and Rodney Rothstein. 2015. "Cell Biology of Mitotic Recombination." *Cold Spring Harbor Perspectives in Biology* 7 (3): a016535. https://doi.org/10.1101/cshperspect.a016535.
Lowe, Keith, Mauricio La Rota, George Hoerster, Craig Hastings, Ning Wang, Mark Chamberlin, Emily Wu, Todd Jones, and William Gordon-Kamm. 2018. "Rapid Genotype 'Independent' Zea Mays L. (Maize) Transformation via Direct Somatic Embryogenesis." *In Vitro Cellular & Developmental Biology - Plant* 54 (3): 240–52. https://doi.org/10.1007/s11627-018-9905-2.
Lowe, Keith, Emily Wu, Ning Wang, George Hoerster, Craig Hastings, Myeong-Je Cho, Chris Scelonge, et al. 2016. "Morphogenic Regulators Baby Boom and Wuschel Improve Monocot Transformation." *The Plant Cell* 28 (9): 1998–2015. https://doi.org/10.1105/tpc.16.00124.
Maher, Michael F., Ryan A. Nasti, Macy Vollbrecht, Colby G. Starker, Matthew D. Clark, and Daniel F. Voytas. 2020. "Plant Gene Editing through de Novo Induction of Meristems." *Nature Biotechnology* 38 (1): 84–89. https://doi.org/10.1038/s41587-019-0337-2.
Mali, P., L. Yang, K. M. Esvelt, J. Aach, M. Guell, J. E. DiCarlo, J. E. Norville, and G. M. Church. 2013. "RNA-Guided Human Genome Engineering via

Cas9." *Science* 339 (6121): 823–26. https://doi.org/10.1126/science.1232033.
Mara, Kostlend, Florence Charlot, Anouchka Guyon-Debast, Didier G. Schaefer, Cécile Collonnier, Mathilde Grelon, and Fabien Nogué. 2019. "POLQ Plays a Key Role in the Repair of CRISPR/Cas9-Induced Double-Stranded Breaks in the Moss Physcomitrella Patens." *New Phytologist* 222 (3): 1380–91. https://doi.org/10.1111/nph.15680.
Maresca, Marcello, Victor Guosheng Lin, Ning Guo, and Yi Yang. 2013. "Obligate Ligation-Gated Recombination (ObLiGaRe): Custom-Designed Nuclease-Mediated Targeted Integration through Nonhomologous End Joining." *Genome Research* 23 (3): 539–46. https://doi.org/10.1101/gr.145441.112.
Mookkan, Muruganantham, Kimberly Nelson-Vasilchik, Joel Hague, Zhanyuan J. Zhang, and Albert P. Kausch. 2017. "Selectable Marker Independent Transformation of Recalcitrant Maize Inbred B73 and Sorghum P898012 Mediated by Morphogenic Regulators BABY BOOM and WUSCHEL2." *Plant Cell Reports* 36 (9): 1477–91. https://doi.org/10.1007/s00299-017-2169-1.
Nakade, Shota, Takuya Tsubota, Yuto Sakane, Satoshi Kume, Naoaki Sakamoto, Masanobu Obara, Takaaki Daimon, et al. 2014. "Microhomology-Mediated End-Joining-Dependent Integration of Donor DNA in Cells and Animals Using TALENs and CRISPR/Cas9." *Nature Communications* 5 (1): 1–8. https://doi.org/10.1038/ncomms6560.
Opperman, Roy, Eyal Emmanuel, and Avraham A. Levy. 2004. "The Effect of Sequence Divergence on Recombination Between Direct Repeats in Arabidopsis." *Genetics* 168 (4): 2207–15. https://doi.org/10.1534/genetics.104.032896.
Ordon, Jana, Johannes Gantner, Jan Kemna, Lennart Schwalgun, Maik Reschke, Jana Streubel, Jens Boch, and Johannes Stuttmann. 2017. "Generation of Chromosomal Deletions in Dicotyledonous Plants Employing a User-Friendly Genome Editing Toolkit." *The Plant Journal: For Cell and Molecular Biology* 89 (1): 155–68. https://doi.org/10.1111/tpj.13319.
Orlando, Salvatore J., Yolanda Santiago, Russell C. DeKelver, Yevgeniy Freyvert, Elizabeth A. Boydston, Erica A. Moehle, Vivian M. Choi, et al. 2010. "Zinc-Finger Nuclease-Driven Targeted Integration into Mammalian Genomes Using Donors with Limited Chromosomal Homology." *Nucleic Acids Research* 38 (15): e152–e152. https://doi.org/10.1093/nar/gkq512.
Paszkowski, Jerzy, Markus Baur, Augustyn Bogucki, and Ingo Potrykus. 1988. "Gene Targeting in Plants." *The EMBO Journal* 7 (13): 4021–26. https://doi.org/10.1002/j.1460-2075.1988.tb03295.x.
Phillips, R. L., S. M. Kaeppler, and P. Olhoft. 1994. "Genetic Instability of Plant Tissue Cultures: Breakdown of Normal Controls." *Proceedings of the National Academy of Sciences* 91 (12): 5222–26. https://doi.org/10.1073/pnas.91.12.5222.

Pinto, Ameet J., and Lutgarde Raskin. 2012. “PCR Biases Distort Bacterial and Archaeal Community Structure in Pyrosequencing Datasets.” *PLoS ONE* 7 (8). https://doi.org/10.1371/journal.pone.0043093.

Polz, Martin F., and Colleen M. Cavanaugh. 1998. “Bias in Template-to-Product Ratios in Multitemplate PCR.” *Applied and Environmental Microbiology* 64 (10): 3724–30.

Pomeroy, Emily J., John T. Hunzeker, Mitchell G. Kluesner, Walker S. Lahr, Branden A. Smeester, Margaret R. Crosby, Cara-lin Lonetree, et al. 2020. “A Genetically Engineered Primary Human Natural Killer Cell Platform for Cancer Immunotherapy.” *Molecular Therapy* 28 (1): 52–63. https://doi.org/10.1016/j.ymthe.2019.10.009.

Potapov, Vladimir, and Jennifer L. Ong. 2017. “Examining Sources of Error in PCR by Single-Molecule Sequencing.” *PLOS ONE* 12 (1): e0169774. https://doi.org/10.1371/journal.pone.0169774.

Puchta, H., B. Dujon, and B. Hohn. 1996. “Two Different but Related Mechanisms Are Used in Plants for the Repair of Genomic Double-Strand Breaks by Homologous Recombination.” *Proceedings of the National Academy of Sciences* 93 (10): 5055–60. https://doi.org/10.1073/pnas.93.10.5055.

Puchta, Holger. 1998. “Repair of Genomic Double-Strand Breaks in Somatic Plant Cells by One-Sided Invasion of Homologous Sequences.” *The Plant Journal* 13 (3): 331–39. https://doi.org/10.1046/j.1365-313X.1998.00035.x.

Puchta, Holger, and Friedrich Fauser. 2015. “Double-Strand Break Repair and Its Application to Genome Engineering in Plants.” In *Advances in New Technology for Targeted Modification of Plant Genomes*, edited by Feng Zhang, Holger Puchta, and James G. Thomson, 1–20. New York, NY: Springer. https://doi.org/10.1007/978-1-4939-2556-8_1.

Qi, Yiping, Xiaohong Li, Yong Zhang, Colby G. Starker, Nicholas J. Baltes, Feng Zhang, Jeffry D. Sander, Deepak Reyon, J. Keith Joung, and Daniel F. Voytas. 2013. “Targeted Deletion and Inversion of Tandemly Arrayed Genes in Arabidopsis Thaliana Using Zinc Finger Nucleases.” *G3: Genes, Genomes, Genetics* 3 (10): 1707–15. https://doi.org/10.1534/g3.113.006270.

Rodríguez-Leal, Daniel, Zachary H. Lemmon, Jarrett Man, Madelaine E. Bartlett, and Zachary B. Lippman. 2017. “Engineering Quantitative Trait Variation for Crop Improvement by Genome Editing.” *Cell* 171 (2): 470-480.e8. https://doi.org/10.1016/j.cell.2017.08.030.

Saika, Hiroaki, Akira Oikawa, Fumio Matsuda, Haruko Onodera, Kazuki Saito, and Seiichi Toki. 2011. “Application of Gene Targeting to Designed Mutation Breeding of High-Tryptophan Rice.” *Plant Physiology* 156 (3): 1269–77. https://doi.org/10.1104/pp.111.175778.

Sánchez-León, Susana, Javier Gil-Humanes, Carmen V. Ozuna, María J. Giménez, Carolina Sousa, Daniel F. Voytas, and Francisco Barro. 2018. “Low-Gluten, Nontransgenic Wheat Engineered with CRISPR/Cas9.” *Plant Biotechnology Journal* 16 (4): 902–10. https://doi.org/10.1111/pbi.12837.

Sanford, John C. 1990. “Biolistic Plant Transformation.” *Physiologia Plantarum* 79 (1): 206–9. https://doi.org/10.1111/j.1399-3054.1990.tb05888.x.
Schendel, Robin van, Sophie F. Roerink, Vincent Portegijs, Sander van den Heuvel, and Marcel Tijsterman. 2015. “Polymerase Θ Is a Key Driver of Genome Evolution and of CRISPR/Cas9-Mediated Mutagenesis.” *Nature Communications* 6 (1): 1–8. https://doi.org/10.1038/ncomms8394.
Schiml, Simon, Friedrich Fauser, and Holger Puchta. 2014. “The CRISPR/Cas System Can Be Used as Nuclease for in Planta Gene Targeting and as Paired Nickases for Directed Mutagenesis in Arabidopsis Resulting in Heritable Progeny.” *The Plant Journal* 80 (6): 1139–50. https://doi.org/10.1111/tpj.12704.
Schmidt, Carla, Michael Pacher, and Holger Puchta. 2019a. “DNA Break Repair in Plants and Its Application for Genome Engineering.” In *Transgenic Plants: Methods and Protocols*, edited by Sandeep Kumar, Pierluigi Barone, and Michelle Smith, 237–66. Methods in Molecular Biology. New York, NY: Springer New York. https://doi.org/10.1007/978-1-4939-8778-8_17.
———. 2019b. “Efficient Induction of Heritable Inversions in Plant Genomes Using the CRISPR/Cas System.” *The Plant Journal* 98 (4): 577–89. https://doi.org/10.1111/tpj.14322.
Shaked, H., C. Melamed-Bessudo, and A. A. Levy. 2005. “High-Frequency Gene Targeting in Arabidopsis Plants Expressing the Yeast RAD54 Gene.” *Proceedings of the National Academy of Sciences* 102 (34): 12265–69. https://doi.org/10.1073/pnas.0502601102.
Shan, Qiwei, Yanpeng Wang, Kunling Chen, Zhen Liang, Jun Li, Yi Zhang, Kang Zhang, et al. 2013. “Rapid and Efficient Gene Modification in Rice and Brachypodium Using TALENs.” *Molecular Plant* 6 (4): 1365–68. https://doi.org/10.1093/mp/sss162.
Shan, Qiwei, Yanpeng Wang, Jun Li, and Caixia Gao. 2014. “Genome Editing in Rice and Wheat Using the CRISPR/Cas System.” *Nature Protocols* 9 (10): 2395–2410. https://doi.org/10.1038/nprot.2014.157.
Silverman, Justin D., Rachael J. Bloom, Sharon Jiang, Heather K. Durand, Sayan Mukherjee, and Lawrence A. David. 2019. “Measuring and Mitigating PCR Bias in Microbiome Data.” *BioRxiv*, April, 604025. https://doi.org/10.1101/604025.
Simmons, Christopher W., Jean S. VanderGheynst, and Shrinivasa K. Upadhyaya. 2009. “A Model of Agrobacterium Tumefaciens Vacuum Infiltration into Harvested Leaf Tissue and Subsequent in Planta Transgene Transient Expression.” *Biotechnology and Bioengineering* 102 (3): 965–70. https://doi.org/10.1002/bit.22118.
Slatko, Barton E., Andrew F. Gardner, and Frederick M. Ausubel. 2018. “Overview of Next-Generation Sequencing Technologies.” *Current Protocols in Molecular Biology* 122 (1): e59. https://doi.org/10.1002/cpmb.59.

Sun, Yongwei, Xin Zhang, Chuanyin Wu, Yubing He, Youzhi Ma, Han Hou, Xiuping Guo, Wenming Du, Yunde Zhao, and Lanqin Xia. 2016. “Engineering Herbicide-Resistant Rice Plants through CRISPR/Cas9-Mediated Homologous Recombination of Acetolactate Synthase.” *Molecular Plant* 9 (4): 628–31. https://doi.org/10.1016/j.molp.2016.01.001.

Suzuki, Keiichiro, Yuji Tsunekawa, Reyna Hernandez-Benitez, Jun Wu, Jie Zhu, Euiseok J. Kim, Fumiyuki Hatanaka, et al. 2016. “*In Vivo* Genome Editing via CRISPR/Cas9 Mediated Homology-Independent Targeted Integration.” *Nature* 540 (7631): 144–49. https://doi.org/10.1038/nature20565.

Suzuki, M. T., and S. J. Giovannoni. 1996. “Bias Caused by Template Annealing in the Amplification of Mixtures of 16S RRNA Genes by PCR.” *Applied and Environmental Microbiology* 62 (2): 625–30.

Symington, Lorraine S., and Jean Gautier. 2011. “Double-Strand Break End Resection and Repair Pathway Choice.” *Annual Review of Genetics* 45 (1): 247–71. https://doi.org/10.1146/annurev-genet-110410-132435.

Szostak, Jack W., Terry L. Orr-Weaver, Rodney J. Rothstein, and Franklin W. Stahl. 1983. “The Double-Strand-Break Repair Model for Recombination.” *Cell* 33 (1): 25–35. https://doi.org/10.1016/0092-8674(83)90331-8.

Terada, Rie, Hiroko Urawa, Yoshishige Inagaki, Kazuo Tsugane, and Shigeru Iida. 2002. “Efficient Gene Targeting by Homologous Recombination in Rice.” *Nature Biotechnology* 20 (10): 1030–34. https://doi.org/10.1038/nbt737.

Thomas, Kirk R., and Mario R. Capecchi. 1987. “Site-Directed Mutagenesis by Gene Targeting in Mouse Embryo-Derived Stem Cells.” *Cell* 51 (3): 503–12. https://doi.org/10.1016/0092-8674(87)90646-5.

Vasil, Indira K., and Vimla Vasil. 1972. “Totipotency and Embryogenesis in Plant Cell and Tissue Cultures.” *In Vitro* 8 (3): 117–25. https://doi.org/10.1007/BF02619487.

Wang, Mugui, Yuming Lu, José Ramón Botella, Yanfei Mao, Kai Hua, and Jian-kang Zhu. 2017. “Gene Targeting by Homology-Directed Repair in Rice Using a Geminivirus-Based CRISPR/Cas9 System.” *Molecular Plant* 10 (7): 1007–10. https://doi.org/10.1016/j.molp.2017.03.002.

Wright, David A., Jeffrey A. Townsend, Ronnie Joe Winfrey, Phillip A. Irwin, Jyothi Rajagopal, Patricia M. Lonosky, Bradford D. Hall, Michael D. Jondle, and Daniel F. Voytas. 2005. “High-Frequency Homologous Recombination in Plants Mediated by Zinc-Finger Nucleases.” *The Plant Journal* 44 (4): 693–705. https://doi.org/10.1111/j.1365-313X.2005.02551.x.

Wu, Hung-Yi, Kun-Hsiang Liu, Yi-Chieh Wang, Jing-Fen Wu, Wan-Ling Chiu, Chao-Ying Chen, Shu-Hsing Wu, Jen Sheen, and Erh-Min Lai. 2014. “AGROBEST: An Efficient Agrobacterium-Mediated Transient Expression Method for Versatile Gene Function Analyses in Arabidopsis Seedlings.” *Plant Methods* 10 (1): 19. https://doi.org/10.1186/1746-4811-10-19.

Wu, Rui, Miriam Lucke, Yun-ting Jang, Wangsheng Zhu, Efthymia Symeonidi, Congmao Wang, Joffrey Fitz, Wanyan Xi, Rebecca Schwab, and Detlef

Weigel. 2018. “An Efficient CRISPR Vector Toolbox for Engineering Large Deletions in Arabidopsis Thaliana.” *Plant Methods* 14. https://doi.org/10.1186/s13007-018-0330-7.

Xu, Yang, Fangquan Wang, Zhihui Chen, Jun Wang, Wen-Qi Li, Fangjun Fan, Yajun Tao, et al. 2019. “Intron-Targeted Gene Insertion in Rice Using CRISPR/Cas9: A Case Study of the Pi-Ta Gene.” *The Crop Journal*, June. https://doi.org/10.1016/j.cj.2019.03.006.

Zelensky, Alex N., Joost Schimmel, Hanneke Kool, Roland Kanaar, and Marcel Tijsterman. 2017. “Inactivation of Pol θ and C-NHEJ Eliminates off-Target Integration of Exogenous DNA.” *Nature Communications* 8 (1): 1–7. https://doi.org/10.1038/s41467-017-00124-3.

Zhang, Congsheng, Changlin Liu, Jianfeng Weng, Beijiu Cheng, Fang Liu, Xinhai Li, and Chuanxiao Xie. 2017. “Creation of Targeted Inversion Mutations in Plants Using an RNA-Guided Endonuclease.” *The Crop Journal* 5 (1): 83–88. https://doi.org/10.1016/j.cj.2016.08.001.

Zhao, Yongping, Congsheng Zhang, Wenwen Liu, Wei Gao, Changlin Liu, Gaoyuan Song, Wen-Xue Li, et al. 2016. “An Alternative Strategy for Targeted Gene Replacement in Plants Using a Dual-SgRNA/Cas9 Design.” *Scientific Reports* 6 (1): 1–11. https://doi.org/10.1038/srep23890.

Zong, Yuan, Qianna Song, Chao Li, Shuai Jin, Dingbo Zhang, Yanpeng Wang, Jin-Long Qiu, and Caixia Gao. 2018. “Efficient C-to-T Base Editing in Plants Using a Fusion of NCas9 and Human APOBEC3A.” *Nature Biotechnology* 36 (10): 950–53. https://doi.org/10.1038/nbt.4261.

Zsögön, Agustin, Tomáš Čermák, Emmanuel Rezende Naves, Marcela Morato Notini, Kai H Edel, Stefan Weinl, Luciano Freschi, Daniel F Voytas, Jörg Kudla, and Lázaro Eustáquio Pereira Peres. 2018. “De Novo Domestication of Wild Tomato Using Genome Editing.” *Nature Biotechnology* 36 (12): 1211–16. https://doi.org/10.1038/nbt.4272.

Zsögön, Agustin, Tomas Cermak, Dan Voytas, and Lázaro Eustáquio Pereira Peres. 2017. “Genome Editing as a Tool to Achieve the Crop Ideotype and de Novo Domestication of Wild Relatives: Case Study in Tomato.” *Plant Science* 256 (March): 120–30. https://doi.org/10.1016/j.plantsci.2016.12.012.